Sprechtempoabhängige Aussprachevariationen

DISSERTATION

zur Erlangung des akademischen Grades
doctor philosophiae
(Dr. phil.)
im Fach Germanistische Linguistik

eingereicht an der
Philosophische Fakultät II
Humboldt-Universität zu Berlin

von
Herrn M.A. Benjamin Weiss
geboren am 16.02.1977 in Bonn

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Dr. h.c. Christoph Markschies

Dekan der Philosophische Fakultät II:
Prof. Dr. Michael Kämper-van den Boogaart

Gutachter:

1. Prof. Dr. Bernd Pompino-Marschall
2. Priv.-Doz. Dr. Hans Georg Piroth

eingereicht am:	11. Februar 2008
Tag der mündlichen Prüfung:	28. Mai 2008

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Teil I

Theoretische Grundlagen

1 Einleitung

Spontansprachliche Äußerungen variieren in Abhängigkeit vom Sprechtempo. Je höher dieses Tempo ist, desto eher weichen Realisierungen von Wörtern, Morphemen und Phonemen von einer kanonischen Aussprache ab. Diese Variationen betreffen sowohl die Anzahl und Art der wortbildenden Allophone, wenn diese etwa assimiliert oder reduziert werden, als auch ihre jeweilige Realisierung, also ihre feineren akustischen Ausprägungen.

Ziel des Forschungsvorhabens ist es, die Bedeutung von Sprechgeschwindigkeit für die Produktion spontansprachlicher Äußerungen zu erfassen. Sprachmaterial des Standard-Deutschen wird anhand einer kanonischen Form untersucht. Aufgrund der dialogischen Gesprächssituation des untersuchten Korpus sind authentische Tempovariationen für diese Art von Kommunikation vorhanden. Die Fragestellung lautet, inwieweit sich das Sprechtempo für diese alltägliche Kommunikationssituation systematisch auf segmentelle[1] Ereignisse (Elisionen, Assimilationen) und auf spektrale Phon-Eigenschaften (wie etwa Formantfrequenzen) auswirkt. Besondere Aufmerksamkeit gilt der Struktur der Veränderungen zum kanonischen Ideal: Unter welchen Bedingungen treten die Veränderungen auf? Wie wirkt sich Sprechgeschwindigkeit auf Parametervariationen aus? Änderungen in der Pausengestaltung werden nicht untersucht.[2]

Datengrundlage der Untersuchung bildet das Kielkorpus (IPDS, 1995–1997), dessen spontansprachlicher Teil einheitlich aus Material einer bestimmten Gesprächssituation (Terminabsprache) mit ihren individuellen Eigenheiten besteht. Analyseergebnisse aus der bestehenden Transkription und eigenen Messungen werden in Verbindung mit dem jeweiligen lokalen Sprechtempo interpretiert. Berücksichtigung finden dabei weitere Einflussfaktoren wie z. B. Betonung und segmenteller Kontext.

[1] In der vorliegenden Arbeit wird der Begriff „Segment“ häufig verwendet. Damit soll nicht impliziert werden, ein Sprachsignal bestünde aus diskreten Einheiten. Dieser Begriff wird entweder verwendet, um auf abstrakter Ebene symbolische Einheiten zu benennen, wie etwa Phoneme oder Silben, oder um (von Mensch oder Maschine) genau festgelegte Abschnitte eines Sprachsignals zu bezeichnen, die damit Phonen oder Silben zugeordnet werden.

[2] Vergleiche dazu etwa Goldman-Eisler (1968); Butterworth (1980).

In der vorliegenden Arbeit werden, soweit dem Autor bekannt, erstmals Aussprachevariationen des Deutschen innerhalb dieser Kontextfaktoren auf ihre Beziehung zu natürlich entstandenen Tempovariationen analysiert. Der überwiegende Teil der in dieser Arbeit zitierten Untersuchungen beruht vor allem auf Labor- oder Vorlesesprache des Englischen. Die hier erreichten Ergebnisse sollen maßgeblich für sachliche Dialoge sein, die gerade für den angewandten Bereich von Sprachtechnologie von Bedeutung sind. Statt gemittelter globaler Maße oder verschiedenster Silben- und Phondauern wird die „Perzipierte lokale Sprechrate" (PLSR) verwendet. Dadurch werden segmentale Einflüsse auf das Tempo minimiert, ohne den lokalen Charakter dieses Teils der Prosodie zu verlieren. Es handelt sich dabei um ein perzeptiv überprüftes Maß, da deutschsprachige Probanden Tempounterschiede in einer Weise wahrnehmen, die sehr stark mit der PLSR korreliert. Die Skala der PLSR ist metrisch und damit besser nutzbar als uneinheitlich begrenzte Kategorien wie „schnell–langsam".

Es wird überprüft, ob bestimmte Effekte, die mit Tempovariation zusammenfallen, auch innerhalb der Grenzen natürlicher Temposchwankungen auftreten, und welchen Systematiken sie folgen. Beispielsweise scheint tempoabhängige Vokalreduktion nicht obligatorisch zu sein (vgl. Kapitel 3.1.1), sodass sich die Frage stellt, wie sich die Sprecher für die weit verbreitete Situation in der Alltagskommunikation verhalten, die das Kielkorpus repräsentiert. Dadurch bieten die Ergebnisse auch eine geeignete Grundlage für die Materialerstellung von relevanten Perzeptionstests.

Das jeweilige Tempo, in dem Wörter, Silben und Laute geäußert werden, hängt von vielen Faktoren ab. Vor allem ist es Ausdruck der Individualität eines Sprechers und der Situation, in der er sich befindet. So ist gelesene Sprache in der Regel langsamer als spontan gesprochene, und im (informellen) Dialog spricht man schneller als bei einem Vortrag (Faust, 1997). Erklärende Passagen in einer Rede sind langsamer als solche, die Zuhörer emotional mitreißen sollen (Weiss, 2005a).

Aber auch die jeweilige Befindlichkeit der Sprecher spielt hierbei eine große Rolle. Ein kurzer Exkurs soll dies für den Ausdruck von Emotionen beim Sprechen darstellen: Ängstlich gesprochene Sätze werden im Mittel schneller artikuliert als neutrale. Langeweile, Ekel, Trauer und Freude führen dagegen zu langsamerer Sprechgeschwindigkeit. Auch lokal zeigen sich charakteristische Unterschiede: Freudig, langweilig und unter Ekel produzierte Äußerungen zeigen erhöhte Silbendauern, besonders für betonte Silben. Trauer dagegen lässt sich zwar auch

durch langsameres Tempo charakterisieren, hier sind es jedoch vor allem die unbetonten Silben, die stark gelängt werden. Unter Ärger entstandene Sprachaufnahmen weisen zwar längere betonte Silben, aber auch verkürzte unbetonte auf. Interessanterweise sind zwar Sätze der Kategorie **Ärger** insgesamt länger als **neutral** produzierte, aber nur aufgrund der Pausen, während die mittlere Silbendauer kaum von neutraler Sprechweise abweicht (Paeschke, 2003; Kienast, 2002).

Es wird deutlich, dass mit Betonung, Pausierung, Befindlichkeit und Sprechsituation zahlreiche Aspekte auf die Sprechgeschwindigkeit einwirken. Weitere nicht genannte lassen sich dazu zählen, z. B. sozio-linguistische Faktoren, wie etwa Alter, Geschlecht oder Dialekt (Verhoeven et al., 2004). Einige solcher Einflussgrößen lassen sich in Untersuchungen kontrollieren, aber längst nicht alle. Deswegen ist es notwendig, vor der eigentlichen Analyse sprechtempobedingter Aussprachevariationen eben solche Faktoren auf Unterschiede im Tempo zu überprüfen, um sie gegebenenfalls kontrollieren zu können (siehe Kapitel 8). Des Weiteren werden schon im letzten Absatz verschiedene Methoden genannt, um Sprechtempo zu messen. Zu diesen Methoden gehören Segment-, Silben- und Wortdauern oder auch Temporaten, letztere in Form von gemittelten produzierten linguistischen Einheiten pro Zeitintervall. Einige Autoren verwenden zur Berechnung der Geschwindigkeit statt tatsächlich geäußerter sprachlicher Segmente auch kanonische Einheiten (z. B. Koreman, 2005).

Über längere Zeiträume gemittelte Werte beschreiben eine globale Geschwindigkeit, einzelne Segmentdauern oder über kürzere Zeiträume erfasste Raten dagegen lokales Tempo. Bei Angaben zum globalen Tempo können Pausen und Häsitationen als *speaking rate* miteinbezogen (Laver, 1994) oder als *articulation rate* ausgelassen werden (Crystal und House, 1988). Da sich die Einbeziehung von Pausen oder Häsitationen und damit die Verwendung der *speaking rate* bei lokalem Tempo ausschließt, wird im Folgenden nicht zwischen Artikulations- und Sprechgeschwindigkeit unterschieden. Auch wird Tempo und Geschwindigkeit synonym verwendet, Sprechrate natürlich nur dann, wenn tatsächlich Raten und keine Dauern bezeichnet werden. Im folgenden Kapitel wird dargelegt, wie sich Sprechgeschwindigkeit phonetisch konstituiert. Eine genaue Definition für den empirischen Teil erfolgt in Kapitel 6. Bisherige Erkenntnisse über das Zusammenwirken von Tempo und Aussprache werden in Kapitel 3 dargestellt. Die perzeptive Verarbeitung von Sprechgeschwindigkeit wird in Kapitel 4 behandelt.

Untersuchungsgegenstand ist spontansprachliches Material, das zudem authen-

tisch ist – also Bedeutung kommunizieren soll (vgl. Hirst und Bouzon, 2005). Trotz fehlender exakter Definitionen hat sich die Trennung von vorbereiteter, abgelesener Rede auf der einen Seite gegenüber spontan konstruierten Äußerungen im Gespräch etabliert. Auch wenn diese Unterscheidung nicht kategorial, sondern graduell ist, lässt sich Spontansprache durch mehr Versprecher, einfachere und unvollständige bis ungrammatische Syntax, begrenzten Wortschatz, mehr Häsitationen und Planungspausen und undeutlichere Artikulation charakterisieren (Faust, 1997). Trotz dieser Eigenschaften wird Spontansprache von Hörern fast mühelos verstanden. Dem gegenüber stellt sie hohe Anforderungen an Verfahren der automatischen Spracherkennung. Entsprechend erhöht sich bei solchen Automaten die Wortfehlerrate für Spontansprache um gut das Vierfache gegenüber Lesesprache (Lippmann, 1997). Und dabei stellt ein sachlicher Dialog zwischen jeweils zwei Personen – wie hier untersucht – sicherlich keinen Extremfall mündlicher Kommunikation dar.

Im Folgenden werden zentrale Ergebnisse aus der Literatur dargestellt. Die meisten gelten für das amerikanische Englisch (AE). Allerdings kann aufgrund vieler Ähnlichkeiten zum Deutschen davon ausgegangen werden, dass die Ergebnisse in ihrer Qualität häufig auf das Deutsche übertragen werden können. Da besonders Sprachen mit anderem Rhythmus und Silbenstruktur unterschiedlich auf Tempovariation reagieren (vgl. z. B. Solé und Ohala, 1991), wird die untersuchte Sprache bei den Literaturverweisen jeweils mit angegeben. Wegen der Dominanz von Studien zum amerikanischen Englisch wird bei dieser Sprache allerdings darauf verzichtet. Bei der Verwendung von Fachbegriffen wird eine einführende Lektüre vorausgesetzt (bspw. Pompino-Marschall, 1995).

2 Sprechgeschwindigkeit und Dauern

Sprechgeschwindigkeit lässt sich den Suprasegmentalia (Lehiste, 1970) zuordnen und hat hauptsächlich etwas mit Quantität, also mit Dauern von Segmenten zu tun. So ist Tempovariation erst einmal dadurch charakterisiert, dass phonologisch vergleichbare Äußerungen unterschiedlich lange dauern. Beim Sprechen wird dies durch die Anzahl und Dauer von Pausen und Segmenten (z. B. Silben oder Lauten) realisiert. Lange Zeit wurde Variation der globalen Sprechgeschwindigkeit allein auf verändertes Pausenverhalten zurückgeführt (Menzerath und de Lacerda, 1933, im Deutschen; Goldman-Eisler, 1968). Und in der Tat können Anzahl und Länge von Pausen durchaus ein Vielfaches des Einflusses von Segmentlängenvariation an der Gesamtlänge von Äußerungen in Lesesprache ausmachen (Crystal und House, 1988). Doch auch die Phone selbst dehnen oder verkürzen sich (Weitkus, 1931, im Deutschen; Weismer und Fennell, 1985). Häufige Beobachtungen sind etwa 20 – 30% Dauerverkürzung bei schnellem gegenüber normalem Sprechen, allerdings abhängig vom jeweiligen Material und Experiment. Dass die Bedeutung der Pausen lange Zeit überschätzt wurde, lag u. a. an der groben Mittelung der Sprechrate, wenn sie z. B. über mehrere Intonationsphrasen hinweg gemessen wurde und so lokale Änderungen und damit verbundene Effekte auf die Aussprache unberücksichtigt blieben (vgl. Miller et al., 1984).

Bei der vorliegenden Untersuchung zum Einfluss von Sprechgeschwindigkeit auf die Aussprache sind Segmentdauern von besonderem Interesse, da hier die direkte Beziehung zwischen Dauern von sprachlichen Einheiten und ihrer Realisierung untersucht werden kann. Als Aussprache wird in der vorliegenden Arbeit die Realisierung von Phonen und Wörtern verstanden, wodurch Pausensetzung nicht Gegenstand der Untersuchung ist, sondern als Kontext berücksichtigt und als Ausdruck von Sprechtempo in globaler Domäne verstanden wird. So ist die Dauer linguistischer Einheiten (Phone, Silben und auch Wörter) von der Anzahl dieser Einheit innerhalb größerer Domänen abhängig, da sich die Segmente mit höherer Anzahl verkürzen (Nakatani et al., 1981). Für das Kielkorpus ist diese Korrelation für die Domäne zwischen zwei Pausen deutlich höher als für Intonationsphrasen. Diese können in spontaner Sprache einige Pausen bein-

halten (Trouvain et al., 2001, im Deutschen).

2.1 Segmentaler Einfluss auf lokales Tempo

Lokales Tempo lässt sich gut über Silben- und Lautdauern erfassen (vgl. Kapitel 6). Auf lokaler Ebene beeinflusst jedoch phonetischer Kontext das *Timing* – also die Dauern von Silben, Lauten und anderen phonetisch relevanten temporalen Informationen. Deshalb wird in diesem Kapitel der Einfluss segmentalen Kontextes auf das Timing dargestellt. Es sollen hiermit bedeutende Faktoren identifiziert werden, die nicht nur lokale Temposchwankungen konstituieren, sondern gegebenenfalls in dem empirischen Teil zu kontrollieren sind. Nach Veränderungen von Vokal- und Konsonantendauern werden die Auswirkungen der Segmentanzahl auf Dauern von Lauten, Silben und Wörtern dargestellt.

Segmentdauern stellen in der akustischen Phonetik bereits ein breit erforschtes Thema dar. Unbestritten bestimmt die Phonemidentität auch die grundsätzliche Länge eines Phons. Diese Phonem-intrinsische Dauer (Peterson und Lehiste, 1960) scheint dabei artikulatorisch motiviert zu sein. Phoneme weisen je nach Artikulationsbedingungen verschiedene Längen auf, z. B. für Unterschiede in der Artikulationsart. Durchschnittlich sind Frikative länger sind als Plosive (Heid, 1998, im Deutschen) und Vokale kürzer, je höher ihre Zungenlage ist (House, 1961; Dauer, 1981, für das Griechische; O'Shaughnessy 1981, im kanadischen Französisch).

Für gelesene Sprache zeigen Crystal und House (1988) in einer umfassenden Untersuchung, dass Diphthonge länger sind als Monophthonge, gespannte Vokale länger als ungespannte, Sonoranten kürzer als Vokale und Obstruenten kürzer als Vokale und Sonoranten. Dabei sind stimmlose Obstruenten zumindest in betonten Silben länger als stimmhafte (Crystal und House, 1988; Davis und van Summers, 1989). Solche Daueruntersciede zwischen verschiedenen Phonemen erklären in den Daten von Klatt (1975) etwa die Hälfte der beobachteten Varianz in den Segmentdauern.

Aufgrund dieser Unterschiede in der intrinsischen Dauer ergibt sich eine direkte Folge für lokal betrachtete Sprechgeschwindigkeit. Aber die tatsächliche Dauer eines Lautes wird nicht nur von ihrer Identität bestimmt, sondern sie variiert auch in Abhängigkeit ihres jeweiligen Kontextes.

2.1.1 Vokaldauern

House und Fairbanks (1953) untersuchen Vokale in Non-Wörtern mit symmetrischer konsonantischer Umgebung. Die Dauer dieser Vokale wird besonders von dem Merkmal STIMMHAFTIGKEIT[1] der Konsonanten beeinflusst. In stimmhafter Umgebung sind Vokale deutlich länger als in stimmloser. Auch für Artikulationsart und -ort zeigten sich Unterschiede: So sind Vokale kürzer, wenn sie zwischen Plosiven auftreten als zwischen Frikativen und auch zwischen velaren oder bilabialen Konsonanten gegenüber alveolaren.

In späteren Untersuchungen können diese Einflüsse von konsonantischem Kontext grundsätzlich auch für Lesesprache bestätigt werden (Umeda, 1975; Crystal und House, 1988). Für den Einfluss des Artikulationsortes von Konsonanten auf die Vokaldauer ergeben sich allerdings widersprechende Ergebnisse: Bei Peterson und Lehiste (1960), Maak (1953) (im Deutschen) und Fischer-Jørgensen (1964) (für das Dänische) findet sich die Systematik, dass ein Vokal länger ist, je weiter hinten ein nachfolgender Konsonant gebildet wird. Zusätzlich gibt es aber auch gegenteilige Ergebnisse (Luce und Charles-Luce, 1985; Crystal und House, 1988).

Für die systematischen Dauerveränderungen des Vokals ist vor allem der nachfolgende Konsonant verantwortlich (Peterson und Lehiste, 1960; House, 1961; Chen, 1970). Auch wenn der Einfluss prä-vokaler Konsonanten als vernachlässigbar eingestuft wird, ergibt sich zumindest für Vokale nach stimmhaften Plosiven eine signifikante Längung gegenüber stimmlosen (vgl. Allen und Miller, 1999).

Die variierende Vokaldauer vor Konsonanten beeinflusst die Wahrnehmung von STIMMHAFTIGKEIT und ist damit phonemunterscheidend (Denes, 1955; van Santen, 1992). Port und Dalby (1982) zeigen, dass das Dauerverhältnis von Vokal und Verschlussphase des folgenden Konsonanten in silbenfinaler Position ein besseres Korrelat von Stimmhaftigkeit bei Konsonanten als die reine Vokaldauer darstellt. Die Ergebnisse von Luce und Charles-Luce (1985) können diesen Schluss allerdings nicht bestätigen: Unterschiede zwischen stimmhaften und stimmlosen wortfinalen Plosiven werden durch die Dauer vorangehender Vo-

[1] Stimmhaftigkeit wird hier als phonologisches Merkmal betrachtet. Dies bedeutet für germanische Sprachen nicht zwingend ein Auftreten von messbaren Stimmbandschwingungen während der Segmentdauer, sondern kann auch über andere akustische Parameter signalisiert werden. Eine andere typische Bezeichnung ist LENIS–FORTIS, die Kohler (1984b) aufgrund phonetischer Ursachen für treffender erachtet (siehe Lisker, 1982, für eine Kritik). Die genaue Bezeichnung für dieses phonologische Merkmal ist in der vorliegenden Arbeit jedoch nicht von Belang.

kale besser erklärt als durch das Verhältnis von Vokaldauer zu Verschlussdauer (vgl. auch Crystal und House, 1988).

Da dieses Verkürzen von Vokalen vor stimmlosen Obstruenten,[2] wie andere phonetisch/phonologische Einflüsse auf Segmentdauern, in zahlreichen Sprachen auftritt (nach Delattre (1962) im Französischen, Deutschen, Italienischen und Spanischen), wird davon ausgegangen, dass dafür artikulatorische oder rhythmische Zwänge verantwortlich sind und es sich nicht um einen gelernten, also sprachspezifischen Effekt handelt. Allenthalben wird die besondere Stärke in der Ausprägung dieses Effektes im Englischen von Klatt (1976) als Indiz für eine Grammatikalisierung gewertet. Für nicht-germanische Sprachen ist dieser Effekt aber weniger stark ausgeprägt und wird wohl nicht zur Unterscheidung von stimmhaften und stimmlosen Konsonanten verwendet (Chen, 1970; Port et al., 1980; Mack, 1982). In einigen Sprachen tritt er überhaupt nicht auf (Ladefoged und Maddieson, 1996). Dieser Effekt ist in seiner Stärke aber positionsbedingt sehr variabel. Er ist stärker in betonten Silben (Davis und van Summers, 1989), für lange Vokale (Peterson und Lehiste, 1960), vor Frikativen gegenüber Plosiven (House und Fairbanks, 1953), dagegen schwächer, wenn sich der nachfolgende Obstruent in der nächsten Silbe eines Wortes befindet (Davis und van Summers, 1989). Er ist stärker in phrasenfinaler Position ausgeprägt (Cooper und Danly, 1981), sowie in Lesesprache gegenüber spontaner Sprache (Klatt, 1976). Diese Variabilität veranlasste zu einem Experiment (Laeufer, 1992), mit dem die Autorin zeigt, dass bei Berücksichtigung solcher Unterschiede das britische Englisch (BE) und das Französische in ihrem Ausmaß einen vergleichbaren Effekt von Vokalkürzung vor stimmlosen Obstruenten aufweisen. Dazu vergleicht sie gleichlange Vokale in ähnlicher prosodischer Position. Laeufer (1992) argumentiert deshalb für einen generellen phonetischen Effekt und gegen eine sprachspezifische phonologische Regel.

2.1.2 Konsonantendauern

Die verschiedenen Vokale wurden bisher nicht von einander unterschieden, da sie in ihren Dauern weitgehend vergleichbar auf verschiedene Umgebungen reagieren. Für Konsonanten gilt dies nicht. Diese weisen je nach Kontext individuelle Veränderungen auf. Dabei zeigt der Vokalkontext kaum einen Einfluss,

[2] Dieser Effekt kann genauso gut mit einer Vokallängung vor stimmhaften Obstruenten bezeichnet werden. Die wenigen Untersuchungen zum tatsächlichen Mechanismus dieses Effektes sollen hier nicht diskutiert werden.

aber die benachbarten Konsonanten (Umeda, 1977; Crystal und House, 1988). Beispielsweise beschränkt sich der gerade dargestellte Effekt der Vokalverkürzung vor stimmlosen Obstruenten nicht auf Vokaldauern, sondern gilt auch für Nasale in Vokal-Nasal-Obstruent Verbindungen, die – wie die Vokale vor stimmlosen Obstruenten – verkürzt werden (Chen, 1970; Port, 1981; Crystal und House, 1988).

Die gegenseitige Beeinflussung von Konsonanten wirkt über Wortgrenzen hinaus. Beispielsweise verkürzen sich wortinitial /p/ und /t/, wenn sie einem Konsonanten nachfolgen. /ʃ/ dagegen längt sich, während der Effekt bei anderen Konsonanten vom Kontext anhängt (Umeda, 1977). Die genauen stellungsbedingten Effekte für einzelne Konsonanten sollen deshalb an dieser Stelle nicht dargestellt werden. Umeda (1977) fasst jedoch zusammen, dass sich in der Regel benachbarte Konsonanten gegenüber vokalischem Kontext verkürzen. Ausnahmen bilden Paare mit widersprechenden Gesten für einen Artikulator, wie etwa /ʃ/ und /θ/.

Obstruenten in Konsonanten-Clustern verkürzen sich (Haggard, 1973). Umeda (1977) bestätigt diesen Effekt für die wortinitiale Position und für Cluster zu Beginn von betonten Silben, allerdings mit der Ausnahme /str/, für die sich /s/ leicht längt. Dem gegenüber werden Liquide in Verbindung mit stimmlosen Plosiven länger (Klatt, 1973; Umeda, 1977). Insgesamt sind Konsonanten zu Beginn von silbeninitialen Clustern länger als später, obwohl diese Sequenzen der Sonoritätshierarchie entsprechen, also z. B. Obstruent-Sonorant Kombinationen sind (Bell und Hooper, 1978). Dass in diesem Fall die Sonoranten kürzer sind als erwartet, lässt sich dadurch erklären, dass sie mit dem Vokal ko-produziert werden (Fowler, 1980), ihre Dauer also verkürzt ist, da der Vokal früh beginnt.

Klatt (1976) kommt zu dem Schluss, dass einige dieser Daueruntерschiede als primäre Merkmale in der Perzeption verwendet werden, um z. B. ähnliche Phoneme zu unterscheiden. Zu relevanten Unterschieden im Englischen zählt der Autor lange gegenüber kurzen Vokalen, stimmhafte / stimmlose Frikative, betonte / unbetonte Vokale, phrasenfinale / nicht finale Silben, Stimmhaftigkeit von Konsonanten, signalisiert durch die Dauer vorangehender Vokale, An- / Abwesenheit von Satzakzent.[3] Zahlreiche der bisher beschriebenen Effekte können für vorgelesene Sprache nicht generell bestätigt werden (vgl. Crystal und House, 1988; Anderson und Port, 1994). Teilweise treten sie nur in besonderen Positionen auf, wie etwa der Einfluss von Stimmhaftigkeit auf vorangehende Vo-

[3] Die noch nicht aufgeführten Kontexteffekte werden im folgenden Kapitel behandelt.

kale, der nur direkt vor Pausen messbar ist. Andere Effekte wie der Einfluss der Artikulationsart auf vorangehende Vokale oder Unterschiede in der Verschlussdauer von Plosiven durch Stimmhaftigkeit können nicht bestätigt werden. Dies kann durchaus an der Unausgewogenheit der Daten und an den vielen möglichen supra-segmentalen Faktoren liegen, die Lese- und Spontansprache charakterisieren. Außerdem ist solche Sprache weit redundanter als Einzelwörter und Einzelsätze, und damit müssen möglicherweise phonetisch/phonologische Informationen nicht so deutlich produziert werden.

2.1.3 Segmentanzahl

Die Variabilität in den Segmentlängen ist nicht nur von der Art der benachbarten Segmente abhängig, sondern auch von der Anzahl der Phone in der Silbe und dem Wort: Schon Menzerath und de Oleza (1928) finden für das Spanische heraus, dass Laute und Silben generell kürzer ausgesprochen werden, wenn mehr Segmente innerhalb eines – dann natürlich längeren Wortes – auftreten. Dass Silbenkerne kürzer sind, je mehr Konsonanten ihnen in der Silbe nachfolgen, ist ein bedeutender Grund, diese Konsonanten mit dem Silbenkern gegenüber dem Silbenonset als Reim zusammenzufassen (vgl. Hall, 2000). Während die Abhängigkeit der Segmentdauern von ihrer Anzahl unbestritten ist, gibt es auch widersprechende Resultate für die Silbenanzahl. Harris und Umeda (1974) können für Lesesprache kaum Auswirkungen der Silbenanzahl auf die Vokaldauer feststellen, während der Effekt in betonten Füßen von Wörtern, die in Trägersätzen eingebettet produziert wurden, vorhanden sind (vgl. Klatt, 1973; Carlson und Granström, 1986, für das Schwedische). Im Niederländischen ist der Effekt der Silbenanzahl im Wort für unbetonte Silben geringer als für betonte (Nooteboom, 1972).

Für das britische Englisch erweist sich die Segmentanzahl in der *narrow rhythm unit*[4] als Faktor, der am stärksten negativ mit Phondauern korreliert, während die Domäne der Silbe weit geringere Korrelationen aufweist (Hirst und Bouzon, 2005). Unterschiede in Betonung für die relativen Dauerveränderungen ergeben sich hierbei nicht. Bei der Untersuchung akustischer und artikulatorischer Daten von Silben mit einem oder zwei Konsonanten in der Silbenkoda, finden Munhall et al. (1992) Hinweise darauf, dass die Vokalverkürzung vor Konsonant-Clustern über verstärkte Koartikulation – also stärker überlappende Gesten – realisiert

[4]Diese Domäne umfasst eine betonte samt allen folgenden unbetonten Silben bis zum Wortende.

wird.

2.1.4 Zusammenfassung

Während mittlere Segmentdauern oft als Korrelat für Sprechgeschwindigkeit herangezogen werden, kann lokales Sprechtempo außer bei streng kontrolliertem Material nicht direkt an ihnen abgelesen werden. Wie in diesem Unterkapitel deutlich gemacht wurde, dominiert der phonetische Kontext in Art und Anzahl die jeweiligen Längen von Silben und Phonen derartig, dass einzelne Dauern nicht aussagekräftig lokales Tempo repräsentieren. Für eine sinnvolle Erfassung lokalen Tempos ist eine Minimierung oder Kontrolle von Kontexteinflüssen notwendig, was z. B. mit einer lokalen Mittelung von Dauern erreicht werden könnte. Für die Auswirkungen von variierenden Dauerinformationen auf die wahrgenommene Identität von Segmenten, vergleiche Kapitel 2.3.3 und 2.3.4.

2.2 Supra-segmentaler Einfluss auf lokales Tempo

Nicht nur die direkte phonetische Umgebung beeinflusst Segmentdauern, sondern auch die phonetisch/phonologische Umgebung, die über die Domäne von Segmenten hinausgeht. Des Weiteren finden sich Strukturen höherer linguistischer Ebenen in Segmentdauern wieder (vgl. Klatt, 1976).

Für Sprachen wie das Deutsche und Englische, die eine Tendenz zur sogenannten Isochronie betonter Silben aufweisen (Pike, 1945, Abercrombie, 1967, aber auch Lehiste, 1977, Kohler, 1983, Carlson, 1991) lässt sich linguistischer Einfluss auf Phondauern in zwei Bereiche einteilen: Dem phonetischen Material innerhalb der Silbe und der Prosodie mit ihrem Einfluss auf größere Domänen.[5]

Campbell und Isard vertreten ein Elastizitätsmodell für Dauermodellierung des Britischen, das diesen beiden Einflüssen gerecht wird: Die rahmengebenden Silbendauern werden in einem ersten Schritt vorhergesagt, der die Segmentanzahl in der Silbe, Art des Silbenkerns (Kurz- oder Langvokal, Diphthong, Konsonant),

[5]Hier zeigt sich die Bedeutung von Unterschieden zwischen Sprachen durch verschiedene Arten von Isochronie, da z. B. Dauerinformationen für Vokalidentität im Französischen kaum relevant ist (Gottfried und Beddor, 1988), und damit eine mögliche Übertragbarkeit zahlreicher Ergebnisse tempoinduzierter Veränderungen auf diese Sprache erst einmal nicht gegeben ist. Auch die Variabilität in den Realisierungen von Funktionswörtern wird auf die Isochronie betonter Silben zurückgeführt.

Position der Silbe in Fuß und Phrase, sowie Betonung und die Wortart (Inhalts-, Funktionswort[6]) berücksichtigt (Campbell, 1990). Die jeweiligen Phondauern sind von dieser Silbenlänge abhängig. Dabei wird innerhalb der Silbe genau ein Elastizitätswert angenommen, der jede Segmentdauer modifiziert, indem der Mittelwert mit der modifizierten Standardabweichung einer Referenzgruppe dieser Segmente addiert wird. Diese beiden Werte – Mittelwert und Standardabweichung – beschreiben das phonemintrinsische Verhalten, während eine Kompression oder Dekompression mit nur einem Elastizitätswert für die gesamte Silbe gilt (vgl. Campbell und Isard, 1991). Die genauen Werte sind dabei logarithmisiert, da Lautdauern bis auf einige Ausnahmen[7] logarithmische Verteilungen aufweisen (Rosen, 2005). Bei diesem Ansatz wird nicht postuliert, dass verschiedene Segmente im gleichen Maße variabel seien, sondern dass bei bekannter Variabilität ein lineares Maß für die Gesamtlänge der Silbe angenommen werden kann, das befriedigende Ergebnisse erzielt. Variation des globalen Tempos kann bei diesem Modell über einen konstanten Summanden mit allen Elastizitätswerten erreicht werden.

Die logarithmische Verteilung von Lautdauern mit einem Grenzwert für kleine Längen wurde für ein Indiz für artikulatorisch motivierte Minimaldauern von Phonen gehalten (Klatt, 1976). Empirische Modelle (vgl. van Santen, 1992) zur Vorhersage von Segmentdauern basieren auf der Trennung von phonemspezifischen Dauern, etwa einer intrinsischen Dauer und eben dieser Minimallänge, sowie kombinatorischen Regeln, die auf diese Dauern angewendet werden. Dabei werden die jeweiligen Minimaldauern nicht unterschritten (Klatt, 1979). Allerdings zeigt Port (1981), dass diese Grenzwerte je nach Kontext für das gleiche Allophon weit von einander abweichen. Er schließt daraus, dass sich die Werte nicht minimal artikulatorisch produzierbaren Dauern annähern, sondern dass dieses Phänomen ensteht, weil sprachlich sichergestellt werden müsse, dass in allen Kontexten dauerbasierte phonologische Informationen tatsächlich artikuliert werden können.

Im Folgenden werden die bedeutendsten supra-segmentalen Einflüsse auf Segmentdauern und damit auch auf das lokale Tempo dargestellt.

[6] Zu den Funktionswörtern zählen Pronomina, Artikel, Formverben, Präpositionen, Konjunktionen und Adverbien (Kohler, 1995).

[7] betonte Konsonanten und Phonem /f/

2.2.1 Prominenz

Prominente, also prosodisch hervorgehobene Silben, sind deutlich länger als nicht prominente (Fry, 1955; Parmenter und Treviño, 1935; Lieberman, 1959; Delattre, 1962). Diese Realisierung von Wortakzenten betrifft vor allem den Silbenkern und auch Konsonanten im Onset, am wenigsten silbenfinale Konsonanten (Umeda, 1977; Klatt, 1976; Greenberg et al., 2003b). Artikulatorisch erklärt sich diese Längung prominenter Silben durch besonders deutliche Produktion (de Jong, 1995), vermutlich, um die Verständlichkeit sicher zu stellen. Satzakzent wirkt sich auch durch längeren Silbendauern aus, wobei alle Silben eines Wortes betroffen sind (Turk und Sawusch, 1997; und für das Niederländische Sluijter und van Heuven, 1995).

2.2.2 Wort- und Silbengrenzen

Segmentdauern zeigen innerhalb von Wörtern bestimmte Strukturen, die auch bei der Wahrnehmung verwendet werden, um Silben- und Wortgrenzen zu erkennen (Lehiste, 1970). Beispielsweise werden Obstruent-Sonorant Cluster als silbeninitial klassifiziert, wenn der Obstruent länger ist als der Sonorant (siehe auch vorheriges Kapitel). Weist dagegen der Sonorant eine vergleichbare Dauer auf, wird zwischen den beiden Segmenten eine Silben- bzw. Wortgrenze wahrgenommen (Christie, 1977). Vergleichbare Ergebnisse gelten auch für Vokal-Plosiv-Vokal Abfolgen, wo für Minimalpaare (z. B. "freed Annie", "free Danny") eine längere Konsonantendauer dazu führt, dass die Silben- und Wortgrenze vor den variierten Plosiv gesetzt wird (Boucher, 1988; Tuller und Kelso, 1991). Diese Unterschiede in der Wahrnehmung werden darauf zurückgeführt, dass Konsonanten im Onset länger sind als in der Koda (vgl. Maddieson, 1985).

2.2.3 Silbenschnitt

Ein anderes Phänomen betrifft die Unterscheidung von kurzen und langen Vokalen im Deutschen. Es wird hier im thematischen Bezug zur Silbe dargestellt, weswegen es in Kapitel 2.1 ausgelassen wurde. Der phonologische Status dieser Unterscheidung von kurzen und langen Vokalen ist durchaus umstritten. Dies betrifft insbesondere die Zuordnung des distinktiven Merkmals GESPANNT/UNGESPANNT auf die Gruppe der Lang- und Kurzvokale. Die sogenannten Kurzvokale kommen nur in geschlossenen Silben vor. Folgt im Wort nur ein Konsonant

vor dem nächsten Silbenkern, wird dieser Konsonant als ambisilbisch angenommen. Im Gegensatz zu Langvokalen und Diphthongen können Kurzvokale mehr als einen Konsonanten in der Koda aufweisen. Bis auf die Paare /aː/, /a/ und /ɛː/, /ɛ/[8] sollen sich die beiden Gruppen von Vokalen durch ihre kürzere Dauer, zentralisierte Qualität, geringere Muskelspannung und geringeren sub-glottalen Luftdruck auszeichnen. Von diesen Hypothesen konnte jedoch bisher nur bestätigt werden, dass Kurzvokale kürzer und zentralisiert sind und einen höheren ersten Formanten aufweisen (Fischer-Jørgensen, 1990). Der Dauerunterschied ist allerdings nur in betonten Silben deutlich (vgl. Jessen, 1998). Deshalb ist umstritten, ob der Unterschied in muskulärer Spannung – und damit GESPANNTHEIT – tatsächlich das phonologisch relevante Merkmal zur Erfassung dieser Vokalopposition darstellt, die LÄNGE ist es jedenfalls nicht.

Eine andere Erklärung basiert nicht auf dem distinktiven Merkmal GESPANNTHEIT. Ausgehend von Sievers (1901) führt Vennemann (1991) die Vokalopposition auf den *Silbenschnitt* zurück, der gegenüber GESPANNTHEIT ein prosodisches Merkmal darstellt. Kurzvokale weisen demnach einen scharfen Silbenschnitt auf: Der Kontakt von Vokal und Kodakonsonant ist stark, da der Konsonant die Artikulation des Vokals früh unterbricht. In seiner Dynamik erreicht der Vokal kein Maximum, es fehlt also ein *Decrescendo*. Beim sanften Silbenschnitt dagegen kann der Vokal sein Artikulationsziel erreichen, weil die Verbindung zum Konsonanten schwächer ist. Dadurch ist der Vokal länger und weist ein *Decrescendo* in seiner Intensität auf. Aus dem scharfen Silbenschnitt folgt hier direkt die Beschränkung von Kurzvokalen auf geschlossene Silben und eine mögliche Erklärung der Robustheit der Dauern von Kurzvokalen gegenüber Betonung.

Aber auch dieser Ansatz konnte bisher nicht empirisch bewiesen werden (Fischer-Jørgensen und Jørgensen, 1969). Allerdings zeigen jüngere artikulatorische Untersuchungen, dass die Opposition auf die zeitliche Koordination der Vokal- und Konsonantbewegungen und nicht auf die Art der Bewegungen zurückgeführt werden kann (Hoole et al., 1994; Pompino-Marschall et al., 1998).[9] Bei erhöhter Sprechgeschwindigkeit verkürzen sich betonte Langvokale deutlich, während

[8] Wobei manchmal für /ɛː/ auch kein entsprechendes Gegenstück angenommen wird (siehe Becker, 1998, für eine ausführliche Diskussion).

[9] An dieser Stelle sei angemerkt, das Braunschweiler (1997) keine Kompensation des Dauerunterschieds von /a/ und /aː/ bei nachfolgenden Plosiven feststellt. Während beide eine höhere Dauer vor stimmhaften Plosiven gegenüber stimmlosen aufweisen (siehe Kapitel 2.3 für diesen Effekt), ist die Gesamtdauer einer VC-Verbindung (ohne Aspiration) für den Kurzvokal im Deutschen im Gegensatz zum Langvokal nicht von STIMMHAFTIGKEIT beeinflusst. Die Ursache liegt darin, das die Verschlussdauer von der Dauer des Vokals unabhängig ist. Diese fehlende Kompensation kann als Hinweis auf eine prosodischen Ursache interpretiert werden.

die Dauern der Kurzvokale kaum betroffen sind (Hoole et al., 1994). Trotz dieser weiteren Indizien für die Adäquatheit des Silbenschnitts zur Unterscheidung von Kurz- und Langvokalen bleibt dieser Ansatz unbewiesen.

2.2.4 Phrasenlänge und Position innerhalb von Phrasen

Zum Ende von Äußerungen und Intonationsphrasen kommt es zu einer Längung der letzten Silbe (Oller, 1973; Lehiste, 1973; Klatt, 1975; Gaitenby, 1965; Cooper und Danly, 1981). In dieser ist – im Gegensatz zu anderen prosodischen Einflüssen – überproportional die Dauer des Reim betroffen (Campbell und Isard, 1991; Klatt, 1976). Oller (1973) findet jedoch für betonte Silben auch eine Onset-Längung. Das Auftreten dieser Silbenlängung am Ende von Äußerungen ist unabhängig von einer realisierten Pause und kann die Dauer der Silbe bis zur Verdopplung längen. Für die letzten Silben von Äußerungen kommt es laut Klatt (1976) generell zu einer starken Verlangsamung, die nicht auf die letzte Silbe beschränkt ist.[10]

Ein weiterer wichtiger prosodischer Faktor, der Segmentdauern beeinflusst, ist die Position innerhalb von Intonationsphrasen (Carlson und Granström, 1989, im Schwedischen), da es in deren Verlauf zu einer generellen Verlangsamung (*Rallentando*) kommt (Dankovičová, 1999, 1997, im Tschechischen). Die mittlere Silbendauer verkürzt sich für längere Äußerungen (Nakatani et al., 1981). Innerhalb von Äußerungen treten Längungen auch an Enden von grammatischen Konstituenten auf (Paccia-Cooper und Cooper, 1981; Nakatani et al., 1981; Wightman et al., 1992), und zwar umso deutlicher, je höher die Konstituente in der prosodischen Hierarchie steht (Fougeron und Keating, 1997; Byrd und Saltzman, 1998).[11] Dies gilt auch für Frikative im Deutschen nach prosodischen Grenzen, allerdings nur dann, wenn nicht bereits eine Pause diese Grenze signalisiert (Kuzlaa et al., 2007). Vergleiche dazu auch Keating et al. (2003); Fougeron und Keating (1997) und für das Niederländische Cho und McQueen (2005).

[10] Die Terminologie ist hier nicht immer eindeutig, da mit „utterance-final“ die Enden von Äußerungen oder auch Sätzen gemeint sind, mit „phrase-final“ manche Autoren die Intonationsphrase, andere jedoch syntaktische oder auch phonologische Phrasen bezeichnen. Positionsbedingte Effekte werden für diese Phrasen von manchen Autoren im gleichen Ausmaß angenommen (z. B. Cummins, 1999), für andere nicht (z. B. bei Oller, 1973; Klatt, 1976). Für die finale Längung erscheint die Gleichbehandlung von Intonationsphrase und Äußerung sinnvoll, da bei beiden der Reim der letzten Silbe stärker betroffen ist als der Onset, während dies laut Oller (1973); Campbell und Isard (1991) für Phrasen, die niedriger in der prosodischen Hierarchie stehen, nicht der Fall ist.

[11] Für einen kritischen Überblick zur Unterscheidung syntaktischer und prosodischer Einflüsse, siehe Shattuck-Hufnagel und Turk (1996).

2.2.5 Wortfinale Längung

Neben diesen Phänomenen soll es auch wortfinale Längungen geben (Oller, 1973; Umeda, 1975; Nakatani et al., 1981), die etwa den letzten Konsonanten für das Niederländische betreffen sollen (Cutler et al., 1997). Allerdings können Harris und Umeda (1974) diese in Lesesprache nicht nachweisen. Zumindest einige dieser Ergebnisse können sich auch auf phrasenfinale Längungen oder andere Positionseffekte zurückführen lassen (Turk und Shattuck-Hufnagel, 2000). Während Turk und Shattuck-Hufnagel (2000) keine Hinweise auf wortfinale Längungen finden können, wird der Einfluss der Wortgrenze auf wortinitiale Konsonanten von ihnen bestätigt (siehe dazu auch Oller, 1973; Fougeron und Keating, 1997). Insgesamt überwiegen die Ergebnisse, die im Englischen eine deutliche Vokallängung in finaler Position von Intonationsphrasen und Äußerungen gegenüber solchen von Wörtern ausweisen. Längere Dauern an prosodischen Grenzen müssen jedoch nicht auf dieselbe Ursache zurückgehen. Keating (2006) unterscheidet *initial strengthening* von *final weakening*,[12] da Ergebnisse artikulatorischer Untersuchungen darauf hinweisen, dass es in initialen im Gegensatz zu finalen Positionen zu genaueren und extremeren Artikulationsgesten kommt.

2.2.6 Wortart und Informationswert

Weitere linguistische Einflüsse sind mit Ausnahme von Wortarten bisher kaum untersucht worden: Funktionswörter werden insgesamt kürzer realisiert als phonetisch vergleichbare Inhaltswörter (Kaiki et al., 1990, im Japanischen), wobei diese Trennung aber auch Ausdruck syntaktischer Stellung sein kann (Carlson, 1991). Erste Messungen dazu stammen von Umeda (1975, 1977), die zeigen, dass Vokale und initiale Konsonanten in Funktionswörtern deutlich kürzer sind. Genauso führt höherer Informationswert von Wörtern zu Verlängerungen: Durch Kontext vorhersagbare Wörter sind kürzer und – isoliert dargeboten – schwerer zu verstehen (Lieberman, 1963). Bei Inhaltswörtern korreliert die Wortdauer stärker mit dem regressiven *Bigramm*, also der Wahrscheinlichkeit eines Wortes vor einem bestimmten anderen Wort aufzutreten, als mit der reinen Frequenz. Jedoch ist dieser Effekt geringer für niederfrequente Wörter (Bell et al., 2002). Genauso sind bereits aufgetretene Wörter kürzer als ihre erste Erwähnung im Diskurs (Fowler und Housum, 1987). Das letzte Ergebnis ist auch für das hier untersuchte Korpus bestätigt worden: Das lokale Sprechtempo ist für neu auftre-

[12] Eine Ausnahme hiervon sei das Äußerungsende.

tende Inhaltswörter niedriger als für ihre Wiedererwähnungen (Thoden, 2004). Die beiden Faktoren Wortart und Kontext spielen allerdings zusammen: Häufige oder kontextuell wahrscheinliche Inhaltswörter werden kürzer realisiert als phonetisch vergleichbare seltene (Bell et al., 2002). Dagegen sind Funktionswörter, die in der Regel hochfrequent auftreten, kürzer, wenn sie satzintern stehen oder kontextuell wahrscheinlich sind (Bell et al., 2003).

Erickson (2000) kommt gegenüber dem oben vorgestellten Modell von Campbell und Isard in einer Analyse von Vokallängen zu dem Ergebnis, dass zwei verschiedene Populationen von Daten anzunehmen sind, um die gemessenen Dauern besser erklären zu können. Einsilbige Inhaltswörter und betonte Silben verhalten sich leicht anders als einsilbige Funktionswörter und unbetonte Silben, da Daueреffekte von *final lengthening* und Verkürzung von Vokalen durch nachfolgende stimmlose Obstruenten nur die erste Gruppe signifikant beeinflussen. Hier fallen also die Unterscheidung von Funktionswörten gegenüber Inhaltswörtern mit Betonung zusammen. Davon ist auch der Satzakzent betroffen, da der Satzakzent meist in Wörtern mit hohem Informationsgehalt realisiert wird.

2.2.7 Diskussion

Obwohl sich Segmentdauern über Modelle recht zuverlässig vorhersagen lassen, stellt Port (1981) aufgrund der zahlreichen Einflüsse auf diese Dauern die Frage:

> How can timing be an effective source of phonological information when it is subject to such a variety of overlapping distortions? (Port, 1981, S. 262).

Diese Frage basiert darauf, dass die hier beschriebenen temporalen Informationen nicht nur das lokale Tempo beeinflussen, indem sich segmentale und prosodische Strukturen in den Segmentdauern wiederfinden. Diese Einflüsse verändern auch temporale Informationen, die für die Perzeption genutzt werden, wie *VOT* und Grenzeffekte. Damit bleibt unklar, wie Hörer die temporalen Informationen trotz ihrer Variabilität noch auswerten können. Insbesondere die Reduzierung von Dauerkontrasten durch die Kombination von Kontexteffekten (Klatt, 1973), die, wie im nächsten Kapitel gezeigt wird, durch Tempoveränderungen noch verstärkt werden kann, stellt dabei ein Problem dar. Zusätzlich ergeben sich für diese Kontexteffekte ein unregelmäßiges Auftreten, das sowohl Abhängigkeiten vom Sprechstil als auch vom Individuum zeigt (Crystal und House, 1988; Smith, 2000).

Einen Hinweis zur Lösung des beschriebenen Problems können temporale Parameter darstellen, die trotz der vielen Schwankungen in den Segmentdauern weitgehend invariant bleiben. Im ersten Experiment von Port (1981) werden drei Parameter variiert: Die Anzahl der Silben im Wort, Vokalidentität und Stimmhaftigkeit des postvokalen Konsonanten. Auffällig ist, dass die relativen Dauerunterschiede zwischen den beiden Vokalen /i/ und /ɪ/ für die verschiedenen Bedingungen weitgehend invariant bleiben, genauso wie das Dauerverhältnis zwischen stimmlosem Konsonant und vorangehendem Vokal. Nach seinem zweiten Experiment, in dem zusätzlich die Sprechgeschwindigkeit variiert wird, generalisiert Port, dass sich phonologisch relevante Effekte mit konstanten Dauerverhältnissen kombinieren, andere dagegen nicht, wie solche, die von Silbenanzahl oder Sprechtempo verursacht werden. Dem gegenüber stehen aber Ergebnisse, dass temporale Informationen für betonte und fokussierte Silben ausgeprägter sind als für unbetonte (de Jong, 2004). Beispielsweise ist in unbetonten Silben der Effekt von Stimmhaftigkeit eines Plosivs auf die Länge des vorangehenden Vokals nicht immer signifikant. Des Weiteren beeinflusst die Stimmhaftigkeit des Plosivs seine Verschlussdauer nur für einige Sprecher (Davis und van Summers, 1989). Dieses Thema wird im folgenden Kapitel weiter behandelt, da hier der Einfluss des Sprechtempos eine wesentliche Rolle spielt.

Segmentale und prosodische Einflussfaktoren auf Segmentdauern können bzw. müssen teilweise gleichzeitig auftreten und addieren sich in diesen Fällen. Allerdings erfolgt diese Addition nicht mit absoluten Werten, sondern in der Form, dass bei mehreren Einflüssen die einzelnen Effekte geringer werden. So erklärt sich die Beobachtung minimaler Dauern (siehe Kapitel 2.2).[13]

Insgesamt müssen diese segmentalen und prosodischen Einflüsse auf Laut-, Silben- und Wortdauern kontrolliert oder ausgeschlossen werden. Nur so kann gewährleistet werden, dass zum einen Aussprachevariation tatsächlich mit Sprechgeschwindigkeit korreliert und nicht mit den prosodischen Bedingungen interagiert, so wie beispielsweise Betonung bei Vokalen zu extremeren Formanten längeren Dauern führt (vgl. Kapitel 3.1.1). Zum anderen sollen tempobedingte Unterschiede erfasst werden, auch wenn ihre Systematik für die verschiedenen Bedingungen unterschiedlich ist.

[13] Im Dauermodell von (Klatt, 1976) gibt es mit Ausnahme phrasenfinaler Konsonanten nur Verkürzungsregeln.

2.3 Sprechgeschwindigkeit und temporale Informationen

Im letzten Kapitel wurden bedeutende kontextuelle Einflüsse auf Segmentdauern identifiziert und die sich daraus ergebenden temporalen Informationen zur Phonemunterscheidung beschrieben. Nicht alle Ergebnisse sind widerspruchsfrei. In diesem Kapitel geht es nun erstmalig um Aussprachevariationen, die sich auf die Variabilität von Sprechgeschwindigkeit zurückführen lassen. Es wird dargestellt, wie globales Tempo Segmentdauern verändert und inwiefern temporale Informationen, die für die Sprachverarbeitung relevant sind – beispielsweise die *voice onset time* oder Konsonant-Vokal-Dauerverhältnisse für das Merkmal Stimmhaftigkeit – davon betroffen sind. Da in diesem Kapitel weder spektrale, sondern ausschließlich temporale Effekte diskutiert werden, noch die hier behandelten Parameter Untersuchungsgegenstand im empirischen Teil sind, werden diese Effekte separat von den sprechtempobedingten Aussprachevariation in Kapitel 3 dargestellt. Vielmehr soll gezeigt werden, dass für temporale Informationen bereits umfangreiche Ergebnisse vorliegen, die den Einfluss von globaler Sprechgeschwindigkeit auf Segmentdauern – und damit auch lokales Tempo, auf dauerbasierte Informationen, sowie auf die Wahrnehmung solcherart variierender Sprachrealisierung zeigen.

2.3.1 Phon- und Silbendauern

In einer frühen Untersuchung zeigt bereits Weitkus (1931) für das Deutsche, dass sich Lautdauern verringern, wenn das globale Tempo ansteigt. Doch während er feststellt, dass sich bei höherem Tempo Konsonanten, außer stimmlosen Frikativen, relativ stärker verkürzen als Vokale, weisen neuere Untersuchungen darauf hin, dass wohl das Gegenteil zutrifft: Für das Russische kommen Kozhevnikov und Christovich (1965) zu dem Ergebnis, dass sich die Silbendauern im Wort und die Wortdauern in der Phrase gleichmäßig mit globalem Tempo verändern, ihre relativen Dauern zueinander also über verschiedene Tempi hinweg konstant bleiben. Innerhalb der Silbe ist dies jedoch nicht der Fall. Konsonanten werden bei ansteigendem Tempo in ihrer Dauer weniger komprimiert als Vokale. Das zuletzt genannte Ergebnis wurde auch für das amerikanische Englisch bestätigt (vgl. Lehiste, 1972; Gay, 1978; Max und Caruso, 1997).

Allerdings zeigen spätere Untersuchungen im Gegensatz zu Kozhevnikov und

Christovich (1965), dass es auch keine Konstanz in den relativen Dauern außerhalb der Silbe gibt. Unbetonte Silben sind relativ stärker von ansteigendem Tempo betroffen als wortbetonte (Janse et al., 2003, im Niederländischen) und satzbetonte (Peterson und Lehiste, 1960; Port, 1981), sodass deren Dauerdifferenz sogar noch größer wird. Zur Vollständigkeit wird hier wiederholt, dass im Deutschen erhöhtes Tempo vor allem Dauern von Langvokalen beeinflusst, aber kaum das von Kurzvokalen (Hoole et al., 1994).

Diese Effekte sind allerdings nicht linear. Bei Verlangsamung von durchschnittlichem Tempo werden Phondauern relativ etwa gleichmäßig gelängt (Port, 1976), wie dies Joos (1948) insgesamt für Tempoveränderungen annimmt. Allerdings gibt es weit weniger Untersuchungen zu Dauerveränderungen beim langsamen als beim schnellen Sprechen. Sollte dieses Ergebnis zutreffen, blieben die die Phonemidentität unterstützenden Dauerverhältnisse bei Verlangsamung gleich. Allerdings zeigen die Messungen von Hertrich und Ackermannn (1995), dass bei langsamen Sätzen im Deutschen satzbetonte Wörter prozentual weniger gelängt wurden als der gesamte Satz.

Bei erhöhtem Tempo ergeben sich nicht konstante Verkürzungen von Segmentdauern: Gespannte Vokale im amerikanischen Englisch werden stärker gekürzt als ungespannte und Unterschiede im Verhältnis der Dauern von Vokal und nachfolgendem Konsonant (C/V) zwischen stimmhaften und -losen Konsonanten verringern sich (Gay, 1978; Port, 1981). Erhöhtes Tempo kann also zur Verringerung der Trennschärfe von temporalen Informationen bezüglich phonologischer Kategorien führen. Dies ist insofern von Bedeutung, als temporale Informationen bei der Sprachverarbeitung verwendet werden, wie ab Kapitel 2.3.3 noch ausgeführt werden wird. Betonte Silben zeigen sich dabei robuster als unbetonte.

2.3.2 Betonung und Tempo

Mit einem einfachen Modell hat Lindblom (1963) den Einfluss von Betonung und Tempovariation auf artikulatorische und akustische Realisierungen gleichgesetzt: Sowohl erhöhte Geschwindigkeit als auch fehlende Betonung würde das *Timing*, in diesem Fall die Ansteuerung der Artikulatoren, so verändern, dass sich Artikulationsgesten verkürzen. Die resultierenden Dauerverkürzungen und der spektrale *target undershoot*[14] seien demnach gleichartig und würden da-

[14]Siehe dazu Kapitel 3.1.1.

mit nur von der Dauer, nicht dem Ursprung des *Timings* abhängen. Von dieser Vorstellung motiviert wurden weitere Experimente durchgeführt, die – wie im folgenden Abschnitt dargestellt – zeigen, dass Betonung und Tempo doch verschiedene Effekte hervorrufen.

Gay (1978) kommt zu dem Ergebnis, dass der Einfluss von globalem Tempo auf Vokaldauern stärker ist als der von Betonung. Tuller et al. (1982) zeigt diesen Effekt für Silbendauern. Dem gegenüber messen Crystal und House (1988) und Fourakis (1991) stärkere Veränderung für Betonung. Obwohl durch diese Untersuchungen feststeht, dass Betonung und Tempo nicht dieselben motorischen Auswirkungen haben, zeigen sich diese Unterschiede vor allem in muskulärer Aktivität und im spektralen Bereich (siehe Kapitel 3.1.1). Die Effekte von Betonung und Tempo kombinieren sich unter anderem so, dass schnelle unbetonte Silben die kürzesten sind. So zeigen Botinis et al. (2002), dass die Wirkung auf Segmentdauern für britisches Englisch, amerikanisches Englisch, Griechisch und Schwedisch (mit Ausnahme schwedischer Konsonanten) bei Betonung robust ist, dagegen Silbenposition im Wort und Fokus als Faktoren vernachlässigbar sind. Insgesamt muss jedoch bezweifelt werden, dass die Realisierung von Betonung und besonders Tempo in ihrer Stärke sinnvoll miteinander verglichen werden können, besonders bei unnatürlichen Aufnahmensituationen wie das Ablesen von Wortlisten.

Mögliche Ursachen für Dauerverkürzungen bei höherem Tempo wurden mit artikulatorischen Experimenten untersucht. Auf Grundlage des *task-dynamic models* (Browman und Goldstein, 1989) ergeben sich zwei Möglichkeiten, eine Tempoerhöhung zu erreichen: Schnellere Ausführung von Artikulationsgesten gegenüber stärkerer Überlappung solcher. Akustische Konsequenz wäre im Rahmen des Target-Modells der Unterschied von wenig bis starkem *target-undershoot*. Shaiman (2001) interpretiert ihre Ergebnisse dahingehend, dass Vokale vornehmlich über schnellere Artikulationsgesten realisiert werden, aber je nach Sprecher zusätzlich eine Überlappung realisiert werden kann. Im Vergleich dazu werden Dauerverkürzungen von Konsonant-Clustern mit Überlappungen von Artikulationsgesten erreicht.

2.3.3 Phonemidentität bei Vokalen

Bei Vokalen bedeuten tempobedingte Dauerveränderungen eine direkte Folge für die Phonemidentität, wenn, wie z. B. im Finnischen, Estnischen (Krull et al.,

2003) oder Japanischen (Hirata, 2004), die Quantität der Vokale selbst phonemunterscheidend ist. Hier wirkt sich Sprechtempo auf die Lautdauern und auch direkt auf die Klassifizierung von Lang- gegenüber Kurzvokalen aus. Dagegen existieren unter anderem im Deutschen, Schwedischen oder Englischen Populationen von Vokalen, die auch als LANG gegenüber KURZ bezeichnet werden können (Peterson und Lehiste, 1960). Diese unterscheiden sich allerdings nicht primär in ihrer Dauer, sondern auch durch ihre Qualität.[15]

In solchen Sprachen ist die Dauer kein primäres Merkmal für die phonemische Identität von Vokalen. Allerdings können sich auch hier mit Dauerveränderungen Unterschiede in der Wahrnehmung ergeben. Dies betrifft allerdings nur Vokale, die sich spektral ähneln. So wird ein synthetischer Stimulus, der spektral uneindeutig ist, bei hoher Dauer als Langvokal, bei niedriger Dauer als Kurzvokal identifiziert (Ainsworth, 1972). Die Klassengrenzen sind jedoch variabel und werden von den Dauern der Umgebung beeinflusst (Johnson und Strange, 1982). Kürzere Dauern vor den Stimuli, also ein höheres Sprechtempo, führen dazu, eher Langvokale wahrzunehmen (Ainsworth, 1974). So hat die relative Vokaldauer einen bedeutenden Einfluss auf die Wahrnehmung von verwechselbaren Vokalen (Strange et al., 1983). Es können Langvokale, die mit hohem globalen Tempo produziert wurden, in Isolation oder normalem Tempo als vergleichbare Kurzvokale wahrgenommen werden, andersherum aber nicht (Verbrugge und Shankweiler, 1977; Verbrugge et al., 1976). Diese Asymmetrie wird dadurch erklärt, dass Vokale, die langsam artikuliert werden, bereits selbst ausreichend Informationen zur Identifikation besitzen und daher hohes Tempo keinen Einfluss ausüben kann, während isolierte Darbietung als langsamer Tempokontext interpretiert wird, in dem der Vokal mit seinen akustischen Eigenschaften als Kurzvokal wahrgenommen wird. Während Verbrugge und Shankweiler Klassifizierungsergebnisse als Temponormalisierung von spektraler Reduktion deuten,[16] erklären sich Verbrugge et al. dieses Verhalten als Normalisierung der Vokaldauer. Der Einfluss von Tempounterschieden beschränkt sich auf Kurz-Lang-Paare, gilt aber auch für synthetische Stimuli, die sich zusätzlich in ihren spektralen Eigenschaften in einem geringen Maß unterscheiden (Gottfried et al., 1990). Insofern kann durchaus von einer Normalisierung temporaler und spektraler Informationen gesprochen werden, allerdings nur bezogen auf Unterschiede zwischen phonologisch langen gegenüber kurzen Minimalpaaren, weswegen diese Effekte auch an dieser Stelle dargestellt werden.

[15] Zum Silbenschnitt im Deutschen siehe Kapitel 2.2.

[16] vgl. Kapitel 3.1.1

2.3.4 Phonemidentität bei Konsonanten

Tempoinduzierte Veränderungen bei Konsonanten sind vor allem im temporalen Bereich untersucht worden, da von wenig Einfluss auf phonemunterscheidende spektrale Eigenschaften ausgegangen wird (vgl. Miller, 1981). Zu den temporalen Parametern zählt beispielsweise die *voice onset time* (*VOT*). Die *VOT* bezeichnet die Dauer vom Beginn der Plosivlösung bis zum Beginn des Stimmeinsatzes und gilt sprachübergreifend als einer der bedeutendsten Parameter für die akustische (Lisker und Abramson, 1964) und perzeptive (Lisker und Abramson, 1970) Unterscheidung stimmhafter und stimmloser Plosive in wortinitialer Position. Sie verringert sich bei höherem globalen Tempo für stimmlose Plosive (Summerfield, 1975; Miller et al., 1986; Diehl et al., 1980). Da stimmhafte Plosive bei Summerfield (1975) durchgängig Stimmbandschwingungen aufweisen, wurde die *VOT* für diese Fälle auf Null gesetzt, was dadurch eine Verringerung des Kontrastes zwischen stimmhaften und stimmlosen Plosiven bezüglich der *VOT* bedeutet. Diese Asymmetrie zeigte sich auch in der Studie von Miller et al. (1986). Die Autoren beschrieben allerdings auch, dass die Variation von *VOT* innerhalb der beiden Gruppen von Phonemrealisierungen (stimmhaft/stimmlos) kleiner wird, wenn sich das Tempo erhöht, was die Trennschärfe ja wieder erhöht.

Diese systematische akustische Veränderung wird von Hörern in Abhängigkeit vom Tempo verarbeitet. Bei synthetisch variierten Stimuli verschiebt sich der Grenzwert der *VOT* zwischen der Identifizierung als stimmhafter oder als entsprechender stimmloser Plosiv (z. B. /p/ und /b/) in seiner Systematik vergleichbar zu den akustischen Messungen aus Produktionsdaten (Summerfield, 1981; Miller et al., 1986; Miller und Volaitis, 1989). In den Experimenten von Summerfield (1981) wird deutlich, auf welche Weise Sprechtempo die Verarbeitung von *VOT* bei britischen Sprechern beeinflussen kann: Sowohl die Dauer der vorangehenden Silben wirkt sich auf die Wahrnehmung aus, als auch die Dauer der Silbe, in der sich der Stimulus befindet. Je näher die Tempoinformationen dem Stimulus sind, desto stärker ist ihre Bedeutung. Denn eine Vergrößerung des temporalen Abstandes zwischen Plosiv und vorangehenden Wörtern durch Verlängerung der ursprünglich 50 ms langen Verschlussphase führt zu einem verringerten Einfluss des Tempos dieser vorangehenden Wörter auf die Wahrnehmung des Plosivs. Deswegen spricht Summerfield von einem Zeitfenster von 100 ms für diesen Effekt vorangehender Tempoinformationen, ab dem sich dieser bis 250 ms stark reduziert, um bis etwa 1 s praktisch wegzufallen.[17]

[17] Wobei bereits ab einer Viertelsekunde die Stille selbst bei langsamer Sprechgeschwindigkeit

Bei dem Satz „Why are you bees/peas?“ ist der Einfluss der Dauer des letzten Wortes „you“ weit stärker auf die Klassifizierung des Plosivs als der vorangehenden Wörter in dem Satz. Für die Silbe, in der sich der Stimulus befindet, erhöht eine Vokalverlängerung die Klassengrenze der *VOT*, was als Verlangsamung der lokalen Geschwindigkeit interpretiert werden kann. Dagegen verringert sich die Kategoriengrenze, wenn der stimmhafte Frikativ /z/ im Kontext „/bi/ again“ an die Zielsilbe angefügt wird, unabhängig von der Dauer dieses Frikatives. Dagegen zeigt die Dauer des nachfolgenden „again“ keinen Einfluss auf die Kategoriengrenze der Stimmhaftigkeit bei der Wahrnehmung. Durch diese verschiedenen Einflüsse ergeben sich schon erste Hinweise darauf, dass die Domäne starker Auswirkungen von Sprechgeschwindigkeit zwar über die Dauer des betreffenden oder direkt benachbarten Segments hinausgeht, aber dennoch lokaler ist als ein Satz oder eine Intonationsphrase. Summerfield (1981) interpretiert seine Ergebnisse wie folgt:

> This observation implies that rate cannot mediate the interpretation of the elements; the event must be self-normalizing. Indeed, to speak of normalization as in an engineering approach to speech recognition is unnecessary, because it appears that timing is largely intrinsic to the acoustical specification of phonetic identity. (S. 1091)

Auf die Verarbeitung tempoinduzierter Variation wird detailliert im Kapitel 6 eingegangen.

Wie komplex der Zusammenhang von Tempo auf *Timing* ist, zeigen die Experimente von Miller und Volaitis (1989) und Wayland et al. (1994). Variierende Silbendauern führten bei Probanden nicht nur zu Verschiebungen der Kategoriengrenzen der *VOT*-Werten von Plosiven, sondern auch zu veränderten Breiten der Kategorieprototypen. So verringert sich für kürzere Silbendauern auch die mögliche Spannbreite der *VOT*-Werte, die einen Stimulus zu einem sehr guten Vertreter seiner Kategorie machen. Für globale Tempovariation bleibt diese Spanne aber gleich (vgl. dazu auch Summerfield, 1981; Kidd, 1989). Dieses Ergebnis wird von Wayland et al. (1994) als Indiz gewertet, dass lokales (Silben-) und globales (Satz-) Tempo unterschieden werden müssen. Zu einem ähnlichen Ergebnis kommen Pompino-Marschall und Janker (1999) in einem anderen thematischen Bereich, nämlich der Perzeption von Silbigkeit: Stimuli des Artikels „einen“ ([ʔa͡ɪn] gegenüber [ʔa͡ɪn̩]) führen bei einer längeren Nasaldauer zu der Wahrnehmung eines Zweisilbers anstatt eines Einsilbers. Dieser Effekt ist unab-

nicht mehr als Verschlussphase, sondern als Einfügen einer Pause interpretiert werden dürfte.

hängig von der lokalen Sprechrate, aber nicht vom globalen Tempo.

Globale Tempovariation hat einen gleichmäßigen Effekt auf die Produktion von *VOT*, Vokal- und relative Konsonantendauer, sodass selbst der C/V-Quotient (siehe unten) ähnlich bleibt (Kessinger und Blumstein, 1998). Die Autoren folgern, dass zahlreiche Experimente zur Wahrnehmung, deren Stimulusmaterial zu Gunsten einer konstanten Silbendauer antiproportionale Dauern von *VOT* und Vokal aufwiesen, nicht repräsentativ für natürliche Sprache seien. Damit seien möglicherweise auch keine Tempoeffekte untersucht worden, weil sich auch die Vokalwahrnehmung geändert haben könnte.

Trotz veränderter *VOT* in verschiedenen Tempi kann Utman (1998) keinen Tempoeffekt auf die Identifikation natürlicher Plosive bei Dauervariation des nachfolgenden Vokals feststellen. Die Autorin sieht bisherige Ergebnisse, in denen Vokaldauer als Korrelat lokalen Tempos variiert wurde, als Resultat des Experimentdesigns, genauer, der Instruktion. Variation in der Dauer direkt nachfolgender Vokale führt also nicht zwingend zu Neustrukturierungen bei der Perzeption, da entsprechende Stimuli in diesem Fall nicht als anderes Phonem, sondern bloß als schlechtere Vertreter desselben Phonems identifiziert werden. Speziell spricht sich Utman gegen Fowler (1980) aus, nach deren Theorie intrinsischen *Timings* sich andere Identifikationsergebnisse ergeben müssten, da dort *Timing*-Informationen als Teil der Phonemdefinition angesehen werden (vgl. Kapitel 4). So bleibt die Frage, ob *Timing*-Veränderungen innerhalb einer Silbe natürlicherweise zu veränderten *VOT*-Werte und zu Kategoriegrenzveränderungen in der Wahrnehmung von initialen Plosiven führen, oder es sich um ein Artefakt der Experimente handelt, kontrovers diskutiert (als Reaktion auf Utman (1998), vgl. Allen und Miller, 1999).

Zahlreiche akustische Parameter unterscheiden sich für stimmlose und stimmhafte Plosive, die intervokalisch im Wort auftreten. Dazu zählen für stimmlose Plosive u. a. längere *VOT*, kürzere F_1-Transitionen und höherer F_1 im Onset des Folgevokals, höhere Grundfrequenz, längere und intensivere Lösungsgeräusche oder das Fehlen von Stimmbandschwingungen in der Verschlussphase (vgl. Lisker, 1978; Edwards, 1981). Als ein perzeptiv bedeutsamer Parameter gilt die Dauer der Verschlussphase (Lisker, 1957). Auch die Länge des vorangehenden Vokals beeinflusst die Klassifikation von Stimmhaftigkeit (Raphael, 1972). Das Dauerverhältnis von Vokal und Obstruent (C/V-Quotient) in silbenfinaler Position wird von Port und Dalby (1982) als bedeutendster temporaler Parameter für die Wahrnehmung von Stimmhaftigkeit bei Obstruenten bezeichnet (vgl. auch

Denes, 1955; Derr und Massaro, 1980; Port, 1981). Allerdings haben Port und Dalby lediglich Plosive in wortmedialer Position untersucht. Ihr Ergebnis lässt sich besser auf Plosive zwischen zwei Vokalen innerhalb eines Wortes begrenzen.

Ein ganz ähnlicher Parameter, nämlich der Quotient von Vokaldauer zur Summe von Vokal- und Plosivverschlussdauer, dient im Deutschen der Unterscheidung stimmhafter (Quotient größer 0,7) und stimmloser Plosive (Quotient kleiner 0,6) in betonten Silben (Kohler, 1977). Mit einem Perzeptionsexperiment wurde gezeigt, dass dieser Parameter auch Probandenurteile erklärt und die Dauerveränderungen diese stärker beeinflussen als Vokaltransitionen und das Auftreten von Stimmbandschwingungen in der Verschlussphase (Kohler, 1979). Dabei merkt Kohler an, dass, wenn vorhanden, der stärkste Faktor die Aspiration bei stimmlosen Plosiven sei. Die Einbeziehung der Vokaldauer in den Divisor führt im Vergleich zum C/V-Quotienten zu einer Normalisierung verschiedener Vokaldauern. Dabei zeigt Braunschweiler (1997) einen Unterschied zwischen dem Langvokal /aː/ und dem Kurzvokal /a/ im Deutschen: In Vokal-Plosiv-Verbindungen mit [aː] und [a] ist die Verschlussdauer nicht vom Vokal abhängig, variiert aber in Abhängigkeit der eigenen Stimmhaftigkeit. Beide Vokale längen sich antizipatorisch vor stimmhaften Plosiven, der Langvokal absolut stärker als der Kurzvokal. Dadurch ist für den Langvokal die Dauer von Vokal und Verschlussphase mit stimmhaftem Plosiv länger als mit stimmlosem, während Vokal-Plosiv-Verbindungen mit /a/ in ihrer Dauer statistisch unabhängig von der Stimmhaftigkeit sind.

Massaro und Cohen (1983a) argumentieren gegen den C/V-Quotienten als Unterscheidungsmerkmal in der Perzeption. Obwohl dieser geringere Variation für verschiedene Tempi aufweist als Vokal- und Obstruentendauern, sei er doch kein invarianter Parameter. Die getrennte Verarbeitung beider Dauern als unabhängige Merkmale zur Stimmhaftigkeit des betroffenen Obstruenten sei wahrscheinlicher und effizienter.

Die Vokaldehnung vor stimmhaften Obstruenten verringert sich, je höher das Tempo wird (Harris und Umeda, 1974; Port, 1981). Dem gegenüber bleibt das C/V-Verhältnis konstanter, vor allem bei Verlangsamung. Bei hohem Tempo verringert sich aber auch in diesem Quotienten der Kontrast zwischen stimmhaften und stimmlosen Plosiven, vor allem aufgrund der Vokaldauer (Port, 1976). Zugleich verstärkt sich der Unterschied im Auftreten von Stimmbandschwingungen in der Verschlussphase,[18] weshalb wohl ein ganzes Bündel von akustischen

[18] Gemeint ist das Auftreten von Unterbrechungen der Stimmbandschwingung.

Merkmalen zur perzeptiven Stimmhaftigkeits-Unterscheidung herangezogen wird. In einem der Experimente von Port und Dalby (1982) wurde sowohl der Trägersatz als auch das Zielwort „rabbit“ in einem hohen und einem niedrigen Tempo produziert, sodass vier verschiedene Bedingungen getestet wurden. Zusätzlich wurde die Dauer der Verschlussphase variiert. Bei der Klassifikation durch Probanden verschiebt sich die Kategoriengrenze zwischen „rabbit“ und „rapid“ für erhöhtes Tempo zu geringerer Verschlussdauer. Bei dieser Untersuchung haben globales Tempo sowie Wortdauer einen signifikanten Effekt und sie addieren sich. Die Bedeutung der Wortdauer ist jedoch größer. Dieselbe Wortdauer führt demnach bei höherem Tempo zur Wahrnehmung von „rapid“ bereits bei niedrigeren Verschlussdauern. Gegenüber der Verschlussdauer sind die Werte des C/V-Quotienten nicht so stark vom Tempo betroffen. Dieser Parameter ist für lokale Tempovariation weitgehend invariant, verändert sich aber in Abhängigkeit vom globalen Tempo, sodass Stimmhaftigkeit für langsame Trägersätze ab etwa 0,75 und für schnelle ab 0,55 für das Testwort wahrgenommen werden. Beide Tempoeffekte zeigten sich als statistisch unabhängig von einander.

Nicht nur Informationen zur Stimmhaftigkeit, sondern auch zur Artikulationsart von Konsonanten sind temporaler Natur. Bei kürzeren Konsonantentransitionen wird ein Plosiv (z. B. /b/), bei längeren ein Approximant (/w/) wahrgenommen (Liberman et al., 1956). Auch dieser Parameter verändert sich mit variierender Sprechgeschwindigkeit (Gay, 1978; Miller und Baer, 1983) und auch hier verschiebt sich die Kategoriengrenze entsprechend (Miller und Wayland, 1993): Bei längeren Segmentdauern vor der betreffenden Silbe verschiebt sich auch die Kategoriengrenze zu längeren Transitionen (Ainsworth, 1973). Mit einer Reihe von Experimenten können Miller und Liberman (1979) zeigen, dass sich dieser Effekt nicht nur für Tempovariation ergibt, die durch die Transitionsdauer signalisiert wird.[19] Er tritt auch für Tempoänderungen ein, die sich aus der Dauer des Vokals in der relevanten Silbe, in geringerem Maße durch die Dauern in der nachfolgenden Silbe, nicht aber durch die reine Silbendauer ergeben. Die Verlängerung der betreffenden Silbe durch einen angefügten Plosiv bringt einen gegenteiligen Effekt. Dieses Verhalten gleicht dem bezüglich der *VOT*, da erhöhtes Tempo sowohl zu systematischen Verschiebungen in den Kategoriengrenzen bei Perzeptionstests führt, als auch den Unterschied in dem temporalen Parameter verringert, der den Phonemstatus signalisiert. Globale und lokale Korrelate von

[19] Sowohl die Dauern als auch das Ausmaß der Transitionen beeinflusst die Identifikation (Schwab et al., 1981).

Sprechtempo führen zu diesem Effekt. Dabei zeigt das Anfügen eines Konsonanten für die Stimmhaftigkeits- und Artikulationsartunterscheidung eine Wirkung, die als Tempoerhöhung interpretiert werden muss.

2.3.5 Systematiken für temporale Informationen

Miller (1981) merkt am Ende ihrer Literaturschau an:

> We have seen that an alteration in speaking rate introduces complex modifications in a number of properties of speech that serve as cues for phonetic contrast and that listerners adjust, at least within certain limits, to that variation when processing the phonetically relevant information. [...] The magnitude of effect varies across cues, as does the case of wether a cue becomes more or less distinctive with a change in speaking rate. (S. 70)

Miller folgert aus den Ergebnissen bisheriger Studien, dass ein Normalisierungsprozess beim Hörer stattfinden muss, um die Variabilität temporaler Informationen zur Phonemidentität auszugleichen. Wie genau dieser vonstatten geht, insbesondere welche Parameter beim Hören zum Einschätzen des Sprechtempos verwendet werden, ist unklar. Der Parameter C/V-Quotient ist eine Möglichkeit, zumindest Daueränderungen durch Tempovariation innerhalb einer Silbe zu erfassen. Gerade bei den Segmentdauern ist problematisch, dass diese in Relation zum Sprechtempo verarbeitet werden müssen, während sie gleichzeitig dieses Tempo signalisieren. Eingehender wird auf die Perzeption von Sprechgeschwindigkeit in Kapitel 4 eingegangen.

Gegenüber den bisher dargestellten Ergebnissen finden Weismer und Fennell (1985) relativ konstante Dauerverhältnisse zwischen normalem und hohem Tempo, ausgenommen in äußerungsfinaler Position. Dieses Ergebnis betrifft Dauerverhältnisse zwischen Vokalen und Konsonanten sowie zwischen betonten und unbetonten Segmenten. Sie folgern, dass zwar absolute Segmentdauern stark variieren, aber erhöhtes Tempo durch relative gleichmäßige Verkürzungen erreicht wird. Äußerungsfinale Segmente wurden nicht analysiert, da in ihrer Argumentation diese Position eine Besonderheit in der Sprachproduktion darstellt und nicht gleichmäßig von Tempovaration betroffen ist (siehe unten). Zwar erkennen sie eine Tendenz, dass diese Systematik bei sehr hohem Tempo aufgrund minimaler Dauern nicht mehr zutrifft, ansonsten seien aber Dauerverhältnisse robust gegenüber Temposchwankungen.

Anhand der bisherigen Dasrtellung wird deutlich, dass zumindest zwei gegensätzliche Sichtweisen in der Literatur vertreten werden. Es gibt die Annahme, dass Dauern, oder die ihr zugrunde liegende *Timing*-Organisation, sehr komplex auf globale Tempovariation reagiert (z. B. Gay, 1978; Gentner, 1987; Nittrouer et al., 1988; Löfquist, 1991; Shaiman et al., 1995), oder eben sehr einfach, also weitgehend über konstante Veränderungen relativer Dauern (z. B. Schmidt, 1975; Tuller und Kelso, 1984; Weismer und Fennell, 1985; Cummins, 1999). Bei letzterer Sichtweise erklären sich signifikante Abweichungen von dieser Systematik aufgrund von artikulatorischen Zwängen, die in der Regel vernachlässigt werden können (Weismer und Fennell, 1985; Cummins, 1999).[20] Letztendlich betrifft das Ergebnis von Weismer und Fennell (1985) nur gewöhnliche und schnelle Aussprache desselben Satzes. Ihre Annahme gilt aber für zahlreiche Effekte nicht: Dazu gehören Reorganisierungen auf der Phrasenebene, die mit erhöhtem Tempo einhergehen (vgl. Trouvain und Grice, 1999) genauso wenig, wie Dauerveränderungen durch Kontext, etwa durch Anzahl von Segmenten in Silbe und Wort (siehe Kapitel 2.1), Betonung (Davis und van Summers, 1989) oder sprecherspezifische Unterschiede (Smith, 2000). Damit sind also eigentlich alle Effekte nicht betroffen, die auf lokaler Ebene selbst Tempo verändern. Temporale Informationen zur Phonemidentität, wie sie in Fowler (1981) zusammengetragen sind, können somit als relativ invariant für Temposchwankungen im vergleichbaren phonetischen Kontext interpretiert werden. Das C/V-Verhältnis als Parameter für Stimmhaftigkeit zwischenvokalischer Plosive entspricht dieser Sichtweise (vgl. auch die Diskussion in Port und Dalby, 1982). Die Ansicht, dass Dauerverhältnisse von Tempovariation weitgehend nicht beeinflusst werden, entspricht auch der Interpretation von Port (1981). In einer weiteren Überprüfung der Hypothese, dass sich Dauern proportional mit Temposchwankungen verändern, berücksichtigen Max und Caruso (1997) auch die Ergebnisse von Kozhevnikov und Christovich (1965). Diese haben eine solche proportionale Dauerveränderung nicht für Phone in der Silbe, wohl aber für Silbe und Wort nachgewiesen (siehe Kapitel 2.3.1). Doch weder auf Silben- noch auf Satzebene können Max und Caruso (1997) konstante Dauerverhältnisse zwischen drei verschiedenen Sprechgeschwindigkeiten nachweisen. Somit ergeben die zahlreichen Untersuchungen das Gesamtbild, dass keine invarianten Dauerverhältnisse über verschiedene Tempi existieren, auch wenn dies nicht bedeutet, dass Sprachproduktion doch über ein solches *Timing*-Modell abläuft.

[20] Für ein Indiz, dass nicht-lineare Veränderungen in Dauerverhältnissen nicht bewusst produziert werden, um für eine bessere Verständlichkeit zu sorgen, siehe Kapitel 2.3.6.

Im Folgenden werden Auswirkungen von Sprechgeschwindigkeit auf die Ausprägungen temporaler Informationen der Prosodie dargestellt.

Wie gerade erwähnt, reagieren äußerungsfinale Segmente nicht gleichmäßig auf Tempovariation, sondern verschwindet dieser Effekt für langsame Tempi (Bell-Berti et al., 1991; Weismer und Ingrisano, 1979). Cummins (1999) weist über eine Reihe von Tempi eine konstante Verlängerung der Silbendauer für Kontrastakzent und *final lengthening* in der Intonationsphrase nach, die sich sogar beide addieren. Nur besonders hohes Tempo scheint einer eigenen Systematik zu folgen, da sich hier die Verlängerungen stark reduzierten. Bemerkenswert ist die Verwendung von Sprechgeschwindigkeit als Skala, nicht als Kategorie von Cummins (1999).

In neueren Experimenten zu solchen *Timing*-Parametern im amerikanischen Englisch wurde die Wirkung von sprecherspezifischen (Smith, 2000) und induzierten Tempounterschieden (Smith, 2002) untersucht. Wieder bestätigt sich die starke Variationsbreite intrinsischer Tempi. Die Auswirkungen beider Experimente sind weitgehend vergleichbar: Bei schnelleren Äußerungen verringerte sich der Effekt der Vokallängung vor stimmhaften Obstruenten, da solche Vokale noch stärker gekürzt werden als vor stimmlosen. Dagegen steigt relativ der Kontrast von *utterance-final lengthening*[21] für höheres Tempo, während gleichzeitig in dieser Position andere phonemunterscheidende Dauerunterschiede (hohe gegenüber tiefen Vokalen, stimmhafte gegenüber stimmlosen Obstruenten) abgeschwächt sind. Trotz dieser Systematiken sind die genauen Ausprägungen allerdings von Sprecher zu Sprecher unterschiedlich.

2.3.6 Diskussion

Insgesamt führt höheres Tempo zu verkürzten Dauern. Es überwiegt das Ergebnis, dass besser komprimierbare Segmente wie Vokale und lange Konsonanten bei ansteigendem Tempo relativ stärker gekürzt werden als andere. Dies führt zu einer Abschwächung des Kontrastes von dauerbasierten Informationen zur Phonemidentität. Insgesamt geht erhöhtes Tempo also mit Informationsreduktion für diesen Bereich einher. Warum gilt dies aber nicht für die temporalen Informationen über finale Positionen und betonten Silben? Dass diese prosodische Stellungen relativ gestärkt werden, während *Timing*-Informationen im lokalen Kontext reduziert werden, kann eine erlernte Strategie für bessere Perzeption

[21] Der Autor nennt es „phrase-final“.

darstellen (vgl. de Jong und Zawaydeh, 1999) oder universale artikulatorische Ursachen haben (vgl. Weismer und Ingrisano, 1979). Des Weiteren weist Smith (2002) darauf hin, dass das Auftreten von charakteristischen *Timing*-Effekten wie *utterance-final lengthening* und Vokallängung vor stimmhaften Obstruenten durchaus nicht obligatorisch ist, auch wenn er diese Effekte weiterhin für charakteristisch für das amerikanische Englisch hält. Obwohl sich Variation in den Ausprägungen dieser dauerbasierten Effekte durchaus über sprecherspezifische Schwankungen in der Sprechgeschwindigkeit erklären lässt, treten immer noch deutliche Kontext- und Sprecherunterschiede auf.

Die Variabilität der Parameter nimmt bei Erhöhung der Sprechgeschwindigkeit zu. Auch dies ist ein Effekt, der sich nicht auf Dauern beschränkt. Es scheint sich hierbei um ein allgemeines Phänomen für motorische Vorgänge zu handeln (Adams et al., 1993). Adams et al. gehen durch ihre artikulatorischen Messungen davon aus, dass Verlangsamen und Beschleunigen über zwei verschiedene motorische Strategien realisiert wird:

> [...] In particular, the control strategy for speech gestures produced at fast speaking rates appears to involve unitary movements that may be predominantly preprogrammed, whereas gestures produced at slow speaking rates consist of multiple submovements that may be influenced by feedback mechanisms. (S. 41)

Darauf aufbauend müsste eine Re-Analyse bestehender Ergebnisse in Bezug auf relative Beschleunigung oder Verlangsamung zeigen, ob sich verschiedene Systematiken in den Segmentdauern, wie weiter oben bereits beschrieben, vereinfachen ließen. Ein weiterer Hinweis ergibt sich aus Pezeptionsexperimenten: Zeitlich komprimierte Sprache mit ihren für ihr Tempo ja nicht typischen Dauerverhältnissen wird schneller verarbeitet als natürlich schnelle Sprache und dieser gegenüber auch bevorzugt (Janse, 2004). Dieses Verhalten interpretiert die Autorin dahingehend, dass die natürlich auftretenden Nicht-Linearitäten eben nicht bewusst für eine bessere Verständlichkeit erzeugt werden, sondern unbewusst durch Zwänge in der Produktion entstehen. Dies mögen artikulatorische Zwänge sein, die eine verstärkte Koartikulation durch Tempo verhindern, wie sie schon als *coarticulatory resistance* (vgl. Bladon und Al-Bamerni, 1976) ohne Einbeziehung von Tempo für bestimmte phonetische Kontexte bekannt sind. Ein anderer Einfluss kann aber auch darin liegen, dass betonte Silben auf kognitiver Ebene genauer definiert sind als unbetonte, und damit robuster gegenüber koartikulatorischen Effekten (vgl. Cho, 2001). Aus Analysen von Artikulationsdaten

wird deutlich, dass verschiedene Einflussfaktoren auf Segmentdauern auch separaten Mechanismen bei der Produktion entspringen. Für Betonung und Tempo wurde dies bereits am Beginn dieses Kapitel dargestellt. Aber auch Tempo und der Einfluss von Konsonant-Clustern in der Koda unterscheiden sich in der Form, dass Verkürzungen in der Koda sich ausschließlich auf verstärkte Überlappung der Gesten zurückführen lassen, während Vokalverkürzung bei Tempoerhöhung hauptsächlich durch schnellere Gesten erreicht wurde (Shaiman, 2001). Van Summers (1987) zeigt, dass sich Kieferbewegungen von unbetonten Vokalen gegenüber betonten durch langsamere und eingeschränkte Bewegungen auszeichnen, wohingegen Vokale, die aufgrund der Stimmlosigkeit des nachfolgenden Konsonanten kürzer sind, keine Unterschiede im Tempo und Ausmaß der Bewegungen aufweisen. Somit sind die Einflüsse auf Segmentdauern bei weitem noch nicht in ihrer genauen Systematik und ihren artikulatorischen Ursprüngen erfasst. Gerade die artikulatorischen Veränderungen sind hier aber von Bedeutung, da sich auf artikulatorischer Ebene kombinatorische Effekte besser erfassen und ihre Auswirkungen auf temporale und spektrale Aspekte der Akustik gemeinsam beschreiben lassen. Schließlich sind akustische gemessene Dauern nicht immer ein gutes Maß zur Abschätzung artikulatorischen Aufwands.

Wie man sieht, sind Dauermessungen und das Wissen um Sprechtempounterschiede durchaus nicht neu. Allerdings wurde Tempo weder einheitlich, und somit vergleichbar, noch lokal definiert. Bei der Untersuchung von Auswirkungen von Tempo auf temporale und spektrale Eigenschaften von Lauten – letztere werden im folgenden Kapitel behandelt – wurden als Tempomaß vor allem Segmentdauer und mittlere Silben- und Phonrate herangezogen. Weitere Instrumentalisierungen sind besonders in der Perzeptionsforschung zu finden, wo einzelne Dauern gezielt variiert werden, um ihren Einfluss auf Probanden zu testen.

Da es eine Wechselwirkung zwischen Geschwindigkeit und Dauern gibt, ist eine Definition von Tempo zwingend notwendig. Globale Raten zu benutzen ist nicht sinnvoll, da Dauern innerhalb von Phrasen variieren. Es ist auch nicht geklärt, wie „intrinsisch“ sprecherspezifisches globales Tempo überhaupt ist, im Vergleich zur individuellen Gestaltung von Äußerungen, die mit der Länge von Intonationsphrasen auch das Tempo stark beeinflussen (Nakatani et al., 1981; Quené, 2005). Lokales Tempo beinhaltet Informationen über Prosodie, Position und phonetisches Material. Lautdauern selbst sind aber wieder vom Tempo abhängig und damit ein nicht zu kontrollierendes Maß für Sprechgeschwindigkeit bei der Analyse von Spontansprache.

Die in diesem Kapitel behandelten Einflüsse von Tempo auf akustische Parameter temporaler Natur, die vom Hörer zur Phonemklassifikation herangezogen werden, bedeuten keine großen Probleme bei der Perzeption. Dass Menschen sich offenbar diesen Veränderungen anpassen, ist eine der wichtigsten Motivationen für diesen Forschungsbereich. Tempobedingte Variation betrifft aber nicht nur temporale Informationen zum Phonemstatus, sondern auch spektrale Informationen, sowie weitere Aspekte der Aussprache.

3 Tempovariation und Aussprache

Im Folgenden werden wichtige Ergebnisse von tempobedingten Aussprachevariationen dargestellt. Von besonderem Interesse sind akustische *Phon*realisierungen, hier die spektralen Eigenschaften, sowie *Phonem*realisierungen, wie sie in Form einer symbolischen Transkription erfasst werden können. Beide Bereiche sind auch Gegenstand im experimentellen Teil, weswegen sie hier in einem Kapitel präsentiert werden. Auf andere Auswirkungen wie etwa die Intonation wird nicht weiter eingegangen (vgl. dazu z. B. Kohler, 1986; Fougeron und Jun, 1998). Letzter Punkt wird die Wortrealisierung sein, da es bei ansteigendem Tempo zu häufigerem Wegfall von Segmenten oder ganzen Silben kommt (Dalby, 1986). Diese Elisionen werden in der Domäne „Wort“ untersucht, da elidierten Phonemen kein lokales Tempo zugeordnet werden kann.

3.1 Erfassung von Aussprachevariationen mittels spektraler Parameter

Mit Abstand am häufigsten wurden Zusammenhänge zwischen spektralen Eigenschaften und Tempo bei Vokalen untersucht. Doch gerade hier ergibt sich kein abschließendes Bild systematischer Veränderungen. Deshalb wird dieses Thema als erstes und besonders ausführlich behandelt. In diesem Abschnitt werden allgemeinere Aspekte diskutiert, die auch für die weiteren Bereiche gelten.

3.1.1 Monophthonge

Maßgebend für die akustische Charakterisierung und Identifikation von Vokalen sind die ersten zwei bis drei Vokalformaten (Pols et al., 1969). In einem einfachen Target-Modell wird angenommen, dass diese Formanten hypothetische Zielwerte haben, die mit einer charakteristischen Vokaltraktform einhergehen und bei sehr deutlicher Sprache erreicht und einige Zeit (*steady state*) eingehalten werden. Die in dieser Phase (Mitte bis Ende des zweiten Drittels der Segmentdauer) gemessenen Formantwerte sind eine der verbreitetsten Formen akustischer Vo-

kalanalyse.

Konsonantischer Kontext verändert vor allem Vokaltransitionen, der vorausgehende Konsonant stärker als der nachfolgende (Schouten und Pols, 1979). Die Transitionen stellen zusätzliche und wichtige Informationen für die korrekte Identifikation von Vokalen dar (Nearey, 1989). Ihr Einfluss ist aber bereits über ein einfaches Modell mit zwei Werten modellierbar (Hillenbrand und Clark, 2001). Vokalischer Kontext beeinflusst Vokalqualität sogar über Silbengrenzen hinweg (Öhman, 1966).

Die Formantwerte während des *steady state* streuen sehr stark, z. B. zwischen Männern, Frauen und Kindern oder einzelnen Sprechern (Peterson und Barney, 1952). Deswegen wurde versucht, invariante Parameter zu finden, um diese Dialektik von recht eindeutigen perzeptiven Kategorien gegenüber starker akustischer Variabilität aufzulösen (vgl. Perkell und Klatt, 1986).[1] Ainsworth (1975) teilt solche Methoden in intrinsische und extrinsische Verfahren ein, bezogen auf ein jeweiliges Vokalsegment. Werden ausschließlich Informationen eines Segments verwendet, handelt es sich demnach um ein (*Vokal-*) *intrinsisches* Verfahren. Zusätzlich bezeichnet Adank (2003) Verfahren, in denen Formantwerte (und die Grundfrequenz) miteinander verrechnet werden, als *Formant-extrinsisch*, und solche, in denen dies nicht getan wird, als *Formant-intrinsisch*. Einfache Methoden tranformieren absolute Formantwerte (in Hz) in eine andere Skala, um psycho-akustische Effekte von Tonhöhenwahrnehmung (Bark-Skala, Log-Skala u. a.) zu berücksichtigen (demnach Vokal-intrinsisch und Formant-intrinsisch). In dem Modell von Syrdal und Gopal (1986) werden solche transformierten Formantwerte (in Bark, vgl. dazu Traunmüller (1989)) verwendet, indem die Differenzen von Formanten (F_3 und F_2, F_2 und F_1), sowie F_1 und der Grundfrequenz (F_0) gebildet werden (Vokal-intrinsisch und Formant-extrinsisch).

Bei Vokal-extrinsischen Verfahren werden zusätzlich Informationen außerhalb des Vokals zum korrekten Identifizieren herangezogen. Eine inzwischen widerlegte Annahme ist, dass dabei auf sprecherspezifisches Wissen zurückgegriffen wird, indem Hörer einen Vokal über den Vokalraum eines Sprechers normalisieren, den sie durch gehörte Eckvokale (im Englischen die Quasi-Kardinalvokale

[1] Im Weiteren werden ausschließlich Normalisierungmethoden genannt, die sich auf Formantfrequenzen sowie die Grundfrequenz beziehen, da diese am einfachsten zu handhaben und weit verbreitet sind. Die Methoden werden hier nur kurz in ihren Prinzipien dargestellt, ohne den exakten Mechanismus für die Vokalklassifizierung darzulegen. Es wird damit die akustische Vokalrepräsentation über das Spektrum genauso ausgeschlossen, wie komplexere Prozesse in der Perzeption.

/i/, /a/, /u/) einschätzen können (Joos, 1948; Gerstman, 1968). Auch, wenn eine gewisse Vertrautheit mit einem Sprecher die Erkennungsleistungen erhöht, sind es jedoch nicht speziell Informationen über Eckvokale, die bei einer Normalisierung herangezogen werden (Verbrugge et al., 1976). Eine weitere Formant-intrinsische Methode zur extrinsischen Normalisierung sprecherspezifischer Vokalräume betrifft die Einbeziehung von (logarithmierten) Formanfrequenzen als Referenzpunkte, die über zahlreiche Äußerungen gemittelt worden sind (Nearey, 1978). Miller (1989) dagegen verwendet die Differenz logarithmierter Werte (was ein Verhältnis der Informationen entspricht) zur Bestimmung von $F_3 - F_2$, $F_2 - F_1$ und F_1 minus einem Referenzwert für F_0, der sich aus einem Mittelwert für den jeweiligen Sprecher ergibt (Vokal-extrinsisch und Formant-extrinsisch). Bis auf einfache Skalen-Transformationen haben alle diese Methoden gemeinsam, dass absolute Formantwerte über andere spektrale Werte relativiert werden, um sprecherspezifische Unterschiede zu vergleichbaren Werten innerhalb eines Vokalphonems zu normalisieren. Auch wenn diese Methoden den Anspruch haben, menschliche Wahrnehmung zu modellieren, kann ihre Güte vor allem darüber beschrieben werden, wie erfolgreich sie zur Klassifikation von Vokalsegmenten eingesetzt werden können. Ein neuerer Vergleich zeigt, dass unter Beibehaltung sozio-linguistischer Unterschiede solche Methoden am effektivsten sind, die Vokal-extrinsische und Formant-intrinsische Informationen verwerten (Adank et al., 2004). Eine detaillierte Besprechung von sprecherspezifischen Unterschieden von Vokalen und Normalisierungsmethoden findet sich bei Johnson (2005).

Trotz signifikanter Verbesserung durch solche Normalisierungen kann davon ausgegangen werden, dass Formantwerte eines Phonems immer noch stark streuen. Da es einem Sprecher unmöglich ist, Äußerungen akustisch exakt zu wiederholen, und jeder Sprecher besondere Eigenheiten besitzt, ist phonetische und akustische Variation unumgänglich. Verstanden wird er dennoch und große Anteile der Variation spiegeln Kontextinformationen wider. Bei der menschlichen Perzeption werden alle zur Verfügung stehenden Informationen benutzt. Ein Sprecher kann also durchaus ungenau artikulieren – der Einfachheit halber also von einer idealisierten Form abweichen. Beim Hören und Verstehen wird phonetischer Kontext, phonologische, syntaktische und semantische Informationen und pragmatischer Kontext benutzt, um die Identität von Wörtern erschließen zu können (Blache und Meunier, 2004).[2] Gänzlich invariante Parameter sind im

[2]Und wie oft wird in einem informellen Gespräch nachgefragt, weil Sätze tatsächlich nicht verstanden wurden?

akustischen Bereich der Phonetik wohl nicht zu finden (vgl. Kapitel 4). Das bedeutet auch, dass wichtige Informationen und solche, die nicht erschlossen werden können, besonders deutlich übermittelt werden. Diese Tatsache stellt sicherlich einen wichtigen Grund für die akustische Robustheit prominenter Silben oder die lange Dauer neu eingeführter Wörter dar.

Eine bedeutende Modellierung dieses Verhaltens ist die Hypo-Hyper-Speech Theorie (Lindblom, 1990). Sie besagt, dass Sprachproduktion effizient an die Situation angepasst wird. Nur wenn notwendig, wird besonders redundant und präzise artikuliert. Für das Target-Modell bei Vokalen bedeutet dies, der Zielstellung dann nahe zu kommen, wenn die Identifikation des Vokals sichergestellt werden muss. Dagegen kann es aufgrund von Ökonomiegründen auch zu stärkerem Abweichen von Zielstellungen kommen. Folgt man diesem Ansatz, kann spektrale Reduktion als Folge dieses *target undershoots* bei jedem Tempo auftreten (wie z. B. bei Nord, 1986), da Hypo-Speech vor allem mit dem Sprechstil zusammenhängt (Moon und Lindblom, 1994). Aber besonders häufig tritt sie bei erhöhtem Tempo auf, weil *target undershoot* und kürzere Dauern als Aufwandminimierung zusammenfallen. Wenn also besondere Genauigkeit unnötig ist, wird man automatisch sowohl schneller als auch undeutlicher sprechen. Auch wenn insgesamt schneller gesprochen wird, weil z. B. ein Sprecher dies grundsätzlich tut, gewinnt eine ökonomische Artikulation an Bedeutung. *Target undershoot* betrifft alle Allophone und ist nicht mit einer phonologischen, stellungsbedingten Schwa-Reduktion gleichzusetzen.

Während Joos (1948) noch annimmt, dass Tempovariation nur zu einer Dauerkompression führt, während die spektralen Eigenschaften gleich bleiben, ist inzwischen bekannt, dass Tempo artikulatorische Umstrukturierung nach sich zieht, die sehr wohl das Spektrum und damit die Formaten verändern. So ist schnelles Sprechen eine Ursache für Zentralisierungen von Vokalformanten (siehe u. a. Lindblom, 1963). Auch Engstrand und Krull (1989) finden Zentralisierungen bei kürzeren Vokalen für Spontansprache des Schwedischen. Wie bereits erklärt, setzt Lindblom (1963) Segmentdauervariation durch Betonung und Tempo gleich.

Es gibt aber Hinweise darauf, dass Dauerveränderungen von Betonung und Sprechtempo verschieden artikulatorisch motiviert sind (vgl. Kapitel 2.3), wobei muskuläre Aktivität bei Tempounterschieden bei weitem nicht so regulär beeinflusst wird, wie durch Betonung. Gerade Sprechtempounterschiede können anscheinend mit verschiedenen Strategien erreicht werden, je nach dem,

wie deutlich ein Sprecher sein möchte (Flege, 1988). Sie reichen von schnelleren Artikulationsbewegungen auf der einen Seite bis zu gleichem Tempo aber nicht ausgeführten Gesten auf der anderen. Damit gehen Dauervariationen und auch spektrale Veränderungen einher, die nicht unbedingt für Tempo und Betonung gleichartig sein müssen (Tuller et al., 1982). So ist wohl zu erklären, dass andere Untersuchungen kaum *target undershoot* für erhöhtes Tempo finden (Gay, 1978), sehr wohl aber für unbetonte gegenüber betonten Vokalen.

Bei Lindblom (1963) wird *target undershoot* durch Assimilation erklärt. Damit ist Koartikulation gemeint, also akustische Annäherung an den konsonantischen Kontext aufgrund überlappender Artikulationsgesten. Obwohl einige Autoren dieses Phänomen als Zentralisierung verstehen (unter anderem Miller, 1981), kann es im Einzelfall auch zur Dezentralisierung kommen (Lindblom, 1963).[3] Die spektrale Richtung des *target undershoots* hänge von dem Zusammenspiel von konsonantischem Lokus (Delattre et al., 1955) und vokalischem *target* ab, seine Stärke von der Distanz dieser beiden Punkte. Durchschnittlich wird aber ein Schrumpfen des gesamten Vokalraums erwartet. Demgegenüber kann Verbrugge und Shankweiler (1977) nur einen minimalen Effekt auf den Vokalraum finden. Damit sind die Ergebnisse zur Beantwortung der Frage, ob Tempo tatsächlich zu Veränderungen führt, also unklar.

Bei Gay (1978) ergibt sich keine systematische Zentralisierung. Trotz Veränderungen in F_1 und F_2 bei fast allen Vokalen (minimal bei z. B. /i/ und /ɔ/), sind die Differenzen doch recht gering. Zudem handelt es sich nicht immer um Zentralisierungen und die Veränderungen sind nicht für alle vier Sprecher einheitlich. Bei betonten gegenüber unbetonten Vokalen sind jedoch beide Formanten signifikant reduziert. Die Formanttransitionen bleiben trotz Tempounterschieden gleich, der F_2-Lokus ist allerdings näher an der Vokalzielstellung, was Gay als früher beginnende Geste bei schnellerer Artikulation interpretiert.

Auch Engstrand (1988) kann bei Konsonant-Vokal-Konsonant Wörtern (CVC) nur einen Reduktionseffekt für unbetonte gegenüber betonten Silben nachweisen, nicht für Silben mit höherem Tempo.

Ähnliche Ergebnisse stammen von Pols und van Son (1990, 1993). Bei ihren Untersuchungen von Aufnahmen eines Sprechers in zwei Geschwindigkeiten ergibt sich kein spektraler Unterschied zwischen beiden Bedingungen. Der Sprecher hatte vermutlich aktiv das Tempo der Transitionen so an die Vokaldauer ange-

[3] Z. B.: F_2 bei /gæg/, F_1 bei /bɪb/.

passt, dass vergleichbare Zielwerte für die Formanten erreicht wurden.

Auch Fourakis (1991) findet keinen Tempoeffekt auf einzelne Vokale, jedoch ein signifikantes Verkleinern des Vokalraums. Unterschiede in den Formantwerten liegen vor allem am konsonantischen Kontext. Die signifikanten Veränderungen für die einzelnen Vokale sind gering und in zwei Fällen ergibt sich sogar eine Dezentralisierung. Fourakis interpretiert seine Ergebnisse:

> Thus it is expected that changes in tempo and stress may not present a major obstacle for the auditory-perceptual theory, or any theory crucially dependent on invariant acoustic information being present in the speech signal. (S. 1826)

Auf Fourakis (1991) Bezug nehmend argumentieren Moon und Lindblom (1994), dass seine Ergebnisse im Einklang mit der Theorie von Assimilation seien, da der CV Kontext /p/, /h/ kaum Koartikulation nach sich ziehen würde.[4] Einige der Ergebnisse seien wohl auf geringe Lokus-Target Distanzen zurückzuführen. In den Daten von Moon und Lindblom (1994), die bewusst eine großen Lokus-Target Distanz aufweisen (/w/-Vokal-/l/), zeigt sich keine Zentralisierung, sehr wohl aber signifikante Assimilation, die von der Vokaldauer und dem Sprechstil abhängig ist. Ein deutlicherer Stil zeichnet sich bei Ihnen durch geringere Assimilationen und raschere Transitionen aus. Wenn Moon und Lindblom davon sprechen, dass das Tempo konstant gehalten wurde, ist das globale Tempo gemeint. Bei dauerinduzierten Assimilationen eines Wortes an derselben Stelle in einem Trägersatz kann die Vokaldauer als Maß für lokales Tempo angesehen werden.

In einer Analyse von Korpora journalistischer Sprache (Nachrichten und redaktionelle Formate aus Funk und Fernsehen) können Gendrot und Adda-Decker (2005) starke Zentralisierungen von gemittelten Formanten bei kürzeren Segmentdauern finden. Eine Dezentralisierung, die von den Autoren als verstärkte Koartikulation interpretiert wird, ergibt sich für /a/ bei alveolarem Kontext aber nur für das französische Material, nicht für die Daten mit deutscher Sprache.

In wenigen Untersuchungen werden mehr als zwei bis drei Kategorien von Sprechgeschwindigkeit oder auch Vokaltransitionen berücksichtigt. Pitermann (2000) ist eine Ausnahme, da er sowohl die Variable **Tempo** metrisch verwendet,

[4] Dieser „Null Kontext" liegt daran, dass sowohl Lippen als auch Glottis nicht unmittelbar mit den für den Vokal wichtigen Artikulatoren Zungenwurzel und -rücken verbunden sind.

als auch Formanttransitionen und -targets analysiert. In seiner Arbeit kommt er zu dem Schluss, dass die Erhöhung der Sprechgeschwindigkeit bei Vokalen weniger zu Zentralisierungseffekten führt, als viel mehr zu Assimilationen zur direkten Umgebung. Trotzdem werden die Transitionen bei ansteigendem Tempo schärfer. Bei dem Unterschied zwischen unbetonten gegenüber betonten Vokalen ist dies jedoch nicht der Fall. Entgegen den Messwerten aus dem Bereich des *steady state* reichen die dynamischen Informationen der Transitionen nicht aus, um die Identifikationsleistungen der Probanden zu erreichen.

Eine neuere Analyse zum Einfluss von Sprechgeschwindigkeit auf Vokalqualität zeigt wiederum keine tempobedingten spektralen Veränderungen für Vokale: Bei Aufnahmen einer Sprecherin sind Formantwerte des *steady state* durch Tempo oder konsonantischen Kontext kaum verändert, obwohl diese Faktoren die Silbendauer systematisch beeinflussen. Stattdessen wird ein deutlicher *target undershoot* gemessen, wenn der Sprechstil von isolierten Wörtern zu ganzen Sätzen wechselt (Stack et al., 2006).

Diese sich widersprechenden Resultate zeigen, dass noch keine Einigkeit darüber herrscht, ob Tempoveränderungen zu Vokalreduktionen führt. Anscheinend kann erhöhtes Tempo Reduktionen auslösen, indem Koartikulation verstärkt wird. Dies kann den Formantraum schrumpfen lassen, da Artikulationsgesten nicht komplett ausgeführt werden. Dies würde unter Berücksichtigung zahlreicher Messwerte erlauben, die Stärke der Reduktion über die Zentralisierung zu messen, da verstärkte Koartikulation weit häufiger zu Zentralisierung als zur Dezentralisierung führt. Je nach Verteilung des phonetischen Kontextes ist dies aber kein verlässliches Maß, wenn die Annäherung nicht in Richtung des Zentrums des Vokalraumes geht (van Bergem, 1995).

Dennoch scheint verstärkte Koartikulation nicht zwingend notwendig zu sein. Wenn gewollt, können Sprecher auch bei erhöhtem Tempo so deutlich artikulieren, dass die *Targets* erreicht werden, indem die Formanttransitionen steiler werden. Handelt es sich also bei tempoinduzierter Vokalreduktion allein um eine Folge des Sprechstils (Moon und Lindblom, 1994)? Da anscheinend zwei verschiedene Strategien angewendet werden können, stellt sich die Frage, welche Strategie im vorliegenden authentischen Material aufftritt und wie robust das Ergebnis ist. Für die hier untersuchten Daten wurde ja kein Sprechstil vorgegeben, sondern enstand natürlich durch die Kommunikationssituation. Gerade Unterschiede zwischen isoliert gesprochenen Wörtern und abgelesenen Sätzen sind für dialogische Kommunikation nicht repräsentativ. Problematisch für ei-

ne These zur Perzeption ist, dass zwei unterschiedliche Produktionsarten eine Normalisierung zu invarianten Parametern ausschließen können. Des Weiteren muss geklärt werden, ob es Unterschiede von den natürlichen Tempovariationen, die Einzelpersonen produzieren, gegenüber denen verschiedener Sprecher gibt. Nicht nur wählen Individuen deutlich unterschiedliche Tempi als „normal“ für sich aus. Auch ihre schnellsten Äußerungen weichen signifikant voneinander ab; und zwar so stark, dass hohe Geschwindigkeiten der intrinsisch langsamen Sprecher etwa dem mittleren Tempo der schnellen Gruppe entsprechen können (Tsao und Weismer, 1997). Während auch Turner et al. (1995) ein tempobedingtes Schrumpfen des Vokalraumes einzelner Sprecher durch generelle Zentralisierungen beobachtet, finden Tsao et al. (2006) keine solchen Effekte für intrinsisch langsame gegenüber schnellen Sprechern. Es gibt auch keine signifikanten Unterschiede für F_1 oder F_2 der einzelnen Vokale zwischen beiden Gruppen. Lediglich die Variabilität im Vokalraum zwischen den langsam sprechenden Personen ist größer. Anders als bei den Untersuchungen zu temporalen Informationen (Kapitel 2.3) ergeben sich hier also mögliche systematische Unterschiede zwischen inter- und intrapersonellen Variationen des Tempos. Eine Bestätigung dieses Ergebnisses wäre ein weiteres Indiz gegen eine einfache Tempo-Normalisierung bei Hören.

3.1.2 Diphthonge

Eine der wenigen Untersuchungen zu den Auswirkungen von höherem Tempo auf Diphthonge zeigt deutliche Zentralisierungen (Gay, 1968). Wichtigstes Ergebnis dieser Arbeit ist aber, dass bei kürzeren Diphthongen die Formanten zu Beginn der Segmente und ihr Verlauf weitgehend gleich bleiben, während sie gegen Ende nicht die Werte vergleichbarer langer Exemplare erreichen, also „abgeschnitten“ sind. Simpson (1998) zeigt für das Kielkorpus ähnliche Effekte, nämlich konstante Transitionsbewegungen, während quasi-stationäre Bereiche besonders zu Beginn der Diphthonge gekürzt und auch leicht zentralisiert werden. Diese Bereiche zu Beginn der Vokale sind für sehr lange Diphthonge nicht stationär, sondern weisen eine Bewegung zu extremeren Formantwerten auf, bevor der eigentliche diphthongale Verlauf einsetzt. Simpson hat ausführliche Messungen durchgeführt, die hier nicht für dasselbe Korpus wiederholt werden sollen.

3.1.3 Konsonanten

Anders als bei Monophthongen wurde in Untersuchungen zu spektralen Eigenschaften z. B. von Obstruenten kaum untersucht, inwieweit Temposchwankungen einen Einfluss auf die Parameter haben, die den Phonemstatus signalisieren. Solche Effekte betreffen bei Konsonanten weitgehend temporale Informationen wie *VOT* und Dauern wie die von Transitionen (siehe Kapitel 2.3). Eine mögliche Ursache für diesen Schwerpunkt auf temporale Informationen wäre, dass – z. B. für Frikative – phonemunterscheidende Auswirkungen fast auszuschließen sind, da sich das Spektrum für einzelne Phoneme doch stark unterscheidet[5] und Transitionen im vokalischen Kontext weitere wichtige Informationen zur Identifikation bereitstellen. Für den Fall von /ʃ/ und /s/ ist jedoch die Identität des nachfolgenden Vokals in der Silbe für unterschiedlichen Interpretationen des Friktionsrauschens desselben Stimulus ausschlaggebend (Fujisaki und Kunisaki, 1978; Repp und Mann, 1980), sodass auch ein Einfluss der Sprechgeschwindigkeit auf die Klassifizierung nicht völlig auszuschließen ist.

Der Fall konsonantischer Reduktion wurde dagegen bereits öfter untersucht. Bei der Artikulation zeigen sich durchaus Reduktionseffekte. Ein Beispiel ist die initiale Stellung in prosodischen Domänen. Hier ergibt sich mehr Gaumenkontakt bei Konsonanten als in anderen Positionen (Fougeron und Keating, 1997). Tempobedingt sind Konsonanten allerdings weniger stark in ihrer Artikulationsbewegung reduziert als Vokale (vgl. Pompino-Marschall, 1995). Im Gegenteil, bei labialen Konsonanten führt die Dauerverkürzung zu schnelleren Bewegungen, die mit stärkerer muskulärer Aktivität einhergehen (Gay und Hirose, 1973). Erklären kann man dieses relative robuste Verhalten dadurch, dass Artikulationsarten wie Plosive und Frikative sich kaum reduzieren lassen, ohne ihre Charakteristika zu verlieren. Schließlich muss für einen tatsächlichen Plosiv zumindest ein kurzer Verschluss produziert werden, um eine ausreichende Engebildung und Strömgeschwindigkeit für ein Friktionsrauschen zu erreichen (Stevens, 1971). Dies betrifft natürlich keine Lenisierung zu einem Frikativ oder Approximant.[6]

Dennoch zeigen artikulatorische Untersuchungen, dass durchaus Phänomene wie Reduktion und verstärkte Koartikulation auftreten, die sich auf Tempoerhöhung zurückführen lassen: Bei höherem Tempo verstärkt sich z. B. die Überlap-

[5] Wenige Frikativpaare lassen sich nicht eindeutig anhand des Rauschspektrums unterscheiden, beispielsweise /f/, /θ/ oder /v/, /ð/ (Harris, 1958).

[6] Der Begriff Lenisierung wird hier nicht als Veränderung des distinktiven Merkmals FORTIS zu LENIS, sondern allgemein als phonetische Abschwächung verstanden.

pung von artikulatorischen Gesten in Form von Zeitdauer zwischen /k/-Burst und Beginn der /l/-Zungenbewegung bei /kl/, unabhängig vom Vokalkontext (Hardcastle, 1985). Verstärkte Koartikulation lingualer Gesten von Konsonanten und Vokalen beschreibt Recasens (1999).

Eine der wenigen akustischen Untersuchungen betrifft Stil und Betonung von informellen Geschichten im Rahmen eines Interviews und den vorgelesenen Transkripten derselben. Hierbei zeigt sich bei Konsonanten außer Plosiven, sowie Vokalen, für spontane Sprache und unbetonte Silben ein vermindertes *center of gravity* (van Son und Pols, 1999). Bei dem *center of gravity* (*COG*), auch *spektrale Balance* genannt, handelt es sich um die über die spektrale Energie gewichtete Mittenfrequenz. Das COG kann als Maß für Artikulationsreduktion bei stimmlosen Konsonanten verwendet werden. Im Fall von Frikativen führt eine Verringerung der Friktionsenge und der Strömgeschwindigkeit der Luft im Mundraum zu einer Reduktion des Rauschens, was sich in einem niedrigeren Wert für das COG niederschlägt. Für stimmhafte Laute hängt es mit dem spektralen Abfall des Anregungssignals zusammen. Damit ist ein niedrigerer Wert die Folge von einem weniger dynamischen Verlauf von Glottalschlägen und charakterisiert damit vor allem den Aufwand bei der Anregung. So ist das COG auch höher für Vokale in betonten Silben (Sluijter und van Heuven, 1996; Sluijter et al., 1997). Für Frauen ist es leicht höher als für Männer. Diese spektralen Unterschiede reichen bei isolierten Frikativen aus, das Geschlecht zu erkennen (Schwartz, 1968).

Auf akustischer Ebene werden besonders zwei Parameter verwendet, um Reduktion und verstärkte Koartikulation zu messen: COG und F_2-Transitionen. Das COG ist ein robustes Maß, da es nicht direkt vom phonetischen Kontext abhängig ist und nicht punktuell, sondern über das ganze Segment erhoben werden kann (mit großen Fensterlänge oder Mittelung mehrerer Werte mit kleiner Fensterlänge bei der Spektralanalyse). Es eignet sich damit besonders für automatisierte Messungen. Dagegen gilt F_2 in seinem Lokus und Bewegung als Korrelat für linguale Koartikulation, da F_2 besonders stark von der Zungenposition beeinflusst wird (Recasens, 1999). Allgemein wird ein niedrigerer F_2 als Zeichen geringeren Tempos der Zungenbewegung gedeutet. F_2 variiert systematisch mit dem Tempo, sodass sich seine Werte z. B. für den Beginn der Verschlussphase von Plosiven gut aus den Werten im zeitlichen Zentrum der umgebenen Vokale und der Verschlussdauer des betroffenen Plosivs vorhersagen lassen. Bei kürzeren Segmenten sinkt der Wert für die Verschlusslösung (Mauk, 2003). F_2 ist aber schwer automatisiert zu messen, da Formantmessungen bei stimmhaften Frika-

tiven durch das Rauschen behindert werden, F_2 bei stimmlosen Obstruenten erst an den Segmentgrenzen auftritt und damit auch zeitlich erkannt werden muss. Zudem gilt es hier, den direkten Kontext mit zu berücksichtigen. Diese praktischen Gründe, zusammen mit der Erkenntnis, dass F_2 ein nicht so reliables Maß darstellt (vgl. van Son und Pols, 1999), führen zu der Entscheidung, im empirischen Teil das COG zu verwenden.

Bei van Son und Pols (1999) handelte es sich um Material von einem professionellen Sprecher. Die Größenordnung der Reduktionen von Konsonanten vergleichen die Autoren mit Ergebnissen von Vokalen und gehen von ähnlichen Auswirkungen auf die Verständlichkeit aus. Akustische Reduktion zeigt sich bei ihnen auch in verringerten Segmentdauern. Ob spektrale und temporale Reduktion auch innerhalb der Kategorien Betonung und Sprechstil korrelieren, wurde nicht untersucht. So führen sie diese Reduktionen auch nicht auf das Tempo zurück, da die Autoren auch bei Vokalen keine tempobedingten Reduktionen nachweisen konnten (vgl. Kapitel 3.1.1). Stattdessen seien Dauerveränderungen mit der spektralen Reduktion eine gemeinsame Folge von Unterschieden im Sprechstil und in dem Merkmal Betonung. In einer neueren Untersuchung von gelesener Sprache zweier Sprecher zeigt sich jedoch ein starker Zusammenhang zwischen Reduktionen von Segmentdauer und COG nicht nur für Betonung, sondern auch für die Position im Wort (Beginn, Mitte, Ende) (van Son und van Santen, 2005).

Hier zeigt sich ein deutlicher Unterschied zum Ansatz in der vorliegenden Arbeit. Van Son und Pols erklären individuelle Tempovariation als Ausdruck struktureller Informationen, die sich dann auch über Kategorien wie Betonung und Position erfassen lassen. Verminderte Dauern werden wie COG-Verringerungen als ein Zeichen von Reduktion behandelt. Es mag durchaus der Fall sein, dass sich gleichzeitige spektrale und temporale Reduktion innerhalb linguistischer Kategorien auf Positionseffekte und damit den Informationsgehalt zurückführen lassen.

Wie aber für Vokale im Kapitel 3.1.1 deutlich gemacht wurde, ist durchaus umstritten, ob für Äußerungen mit gleicher Ausprägung im Sprechstil und in der Promimenz Tempovariation mit spektraler Variation korreliert. Deswegen wird in der vorliegenden Arbeit generell ein solcher Zusammenhang überprüft, und zwar für die Kommunikationsform, in der die Äußerungen produziert wurden. Erst nach einem positiven Ergebnis können mögliche strukturelle Erklärungen innerhalb der verwendeten Kategorien überprüft werden. Der Autor geht jedoch

– bislang implizit – davon aus, dass lokales Tempo selbst, also Segmentdauern von Silbe und Phon, je nach Situation für mögliche spektrale Reduktionen verantwortlich sind, während strukturelle Informationen, wie Segmentposition, artikulatorische Ursachen spektraler Reduktion nur indirekt beeinflussen. Dennoch hätte ein Sprecher durchaus die Möglichkeit, ein positionsbedingt kürzeres Phon mit entsprechendem artikulatorischen Aufwand nicht spektral zu reduzieren. Diese Frage wird in der vorliegenden Arbeit allerdings nicht untersucht. Für Konsonanten im Besonderen stellt sich die Frage, ob sie sich bei Tempovariation grundsätzlich anders verhalten als Vokale, da Konsonanten eindeutigere artikulatorische Zielstellungen aufweisen und z. B. im Koartikulationsmodell von Öhman (1966) als der vokalischen Artikulation überlagert angesehen werden. Damit wird das COG nicht als bedeutender akustischer Parameter für eine phonetisch/phonologische Klassifizierung angesehen, sondern ausschließlich zur Bestimmung der Artikulationsgenauigkeit herangezogen.

3.2 Erfassung von Aussprachevariationen mittels symbolischer Umschrift

Auswirkungen von Tempovariation auf die Aussprache, die sich anhand symbolischer Transkription erfassen lassen, betreffen (soweit bekannt) ausschließlich erhöhtes Tempo. Die Realisierungen von niedrigem oder normalem Tempo werden in der Regel als kanonisch angenommen. Dabei werden die bekannten Veränderungen in verbundener Sprache gegenüber Einzelwörtern und Spontansprache gegenüber Lesesprache als Norm angesehen, anhand derer Abweichungen bei höherem Tempo beschrieben werden.

Ein Effekt von erhöhtem Tempo betrifft mögliche Resilbifizierungen von Wortrealisierungen. Bei der Studie von Laeufer (1995) handelt es sich auch um akustische Analysen. Trotzdem sollen die Ergebnisse in diesem Kapitel vorgestellt werden, da diese akustischen Analysen der Erkennung von Silbengrenzen dienen, die symbolisch dargestellt werden. In den Trägersatz „Dieses Mal habe ich ... gesagt“ wurden verschiedene Wörter eingebettet und von 10 Sprechern langsam und schnell produziert. Realisierte Silbengrenzen wurden von der Autorin allein über akustische Messungen zugewiesen und nicht mit Hilfe von Wahrnehmungsexperimenten bestätigt. Die Analyse zeigt veränderte Silbengrenzen der Zielwörter für schnelles Tempo. Diese tempobedingte Resilbifizierung gilt im Vergleich zu Realisierungen bei langsamer Sprechgeschwindigkeit: Bei

benachbarten unbetonten Silben werden sprachspezifische Restriktionen in der Silbenstruktur aufgegeben, und im Rahmen der Sonoritätshierarchie der Onset der zweiten Silbe verstärkt. So wurden Kodakonsonanten als Onset realisiert, wenn ihre Stärke[7] höher ist als der bisherige Onset ([ɔp.jɛk.ˈtiːf] –> [ɔ.bjɛk.ˈtiːf], aber [an.ti.ˈkvar] bleibt unverändert). Bei vergleichbarer konsonantischer Stärke wird nur manchmal resilbifiziert ([ak.ti.vi.ˈtɛːt] –> [a.kti.vi.ˈtɛt]). Diese Systematik ist laut Laeufer der Trend zu theoretisch optimalen Silben, die sich durch fehlende Koda auszeichnen. Dabei wird die Sonoritäthierarchie nicht verletzt und dem Prinzip des Maximalen Onsets (vgl. Selkirk, 1982) Rechnung getragen. Solche Resibifizierung wird allerdings kaum bei betonten Silben beobachtet. Konsonanten in der Koda unbetonter Silben werden nur bei gleicher konsonantischer Stärke als Onset betonter Silben realisiert ([ɔk.ˈtoːbɐ] –> [ɔ.ˈktoːbɐ]). Für vorangehende betonte Silben ergibt sich ein diversifiziertes Bild. Laeufer (1995) interpretiert ihre Daten dahingehend, dass Silben mit langen Vokalen und Diphthongen nachfolgende initiale Konsonanten eines Clusters in ihre Koda ziehen, was zu zwei verschiedenen Realisierungsmöglichkeiten führt. In seltenen Fällen wird tatsächlich resilbifiziert ([ˈnɔ͡ɪ.trʊm] –> [ˈnɔ͡ɪt.rʊm]). Meist wird jedoch der initiale Konsonant ambisilbisch. Für einzelne Konsonanten ergibt sich eine ambisilbische Realisierung. Nach betonten Silben mit Kurzvokal bleiben einzelne Konsonanten unverändert ambisilbisch. Cluster verhalten sich weitgehend wie bei betonten Silben mit langem Vokal: Für gleiche oder verringernde konsonantische Stärke ergeben sich zwei mögliche Realisierungen, entweder mit der Silbengrenze zwischen beiden Konsonanten oder mit ambisilbischen ersten Konsonanten. Dagegen betrifft bei steigender konsonantischer Stärke die ambisilbische Variante den zweiten der beiden Konsonanten ([ˈsal.ve] –> [ˈsal.ve] oder [ˈsal̩ve]). Die Häufigkeiten dieser beiden Varianten variierten allerdings beträchtlich zwischen verschiedenen Wörtern. Der Vorteil verstärkter betonter Silben sei ein größerer Kontrast zwischen Silbenkern und -grenze. Auch, wenn dies die bisher einzige Studie zu diesem Thema ist, und die Methode der Silbengrenzenbestimmung keine Abschätzung der Reliabilität zulässt, handelt es sich dennoch um eine interessante Untersuchung, deren Ergebnisse auch für Wortübergänge in fließender Rede bedeutsam sein könnten.

Aber nicht nur die Silbifizierung ändert sich mit der Sprechgeschwindigkeit. Je höher das Tempo wird, desto eher treten Elisionen und Assimilationen auf. Weibliche Sprecher weisen bei Funktionswörtern nicht so starke Abweichungen auf

[7]Konsonantische Stärke verhält sich im Rahmen der Sonoritäthierarchie antiproportional zur Sonorität (vgl. Pompino-Marschall, 1995).

wie männliche (Bell et al., 2003). Die Auswirkungen von erhöhtem Tempo werden denen von niedrigerem Register gleichgestellt (vgl. Kohler, 1995). Bei gelesener Sprache wirken sich globale Tempounterschiede auch auf die Realisierung von Wortgrenzen und -übergängen aus. An Wortgrenzen werden bei niedriger Sprechgeschwindigkeit eher Pausen eingefügt und wortfinal Plosivlösungen realisiert als bei normaler. Schnelle Sprechweise fällt durch häufigeres Auftreten von Palatalisierungen und Flapping auf (Moore und Zue, 1985).

Im Deutschen betreffen Elisionen vor allem das Schwa in unbetonten wortfinalen Silben und in *schwachen Formen* (siehe unten). Dagegen wird Schwa nicht getilgt, wenn es einem Obstruent+Nasal nachfolgt. Außerdem sind Geminaten (z. B. /nn/) oder gleiche Plosive (nur in der Koda) und Aspirationen betroffen. Sogenannte *schwache Formen* beinhalten in ihrer Charakterisierung bereits Dauerverkürzungen. Bei häufigen, nicht betonten Funktionswörtern treten alle Arten von Reduktionen (Zentralisierungen, Elisionen) zusammen auf und resultieren in zahlreichen verschiedenen Aussprachevarianten (Kohler, 1990).

Symbolphonetisch annotierte Elisionen bedeuten nicht, dass diese Segmente tatsächlich schon in der Planungsphase wegfallen. Auch wenn sie akustisch kaum oder gar nicht zu erkennen sind, sind zum Teil Reste von lingualen Gesten artikulatorisch vorhanden, was eher auf eine extreme Form von Koartikulation schließen lässt, als auf einen echten Segmentausfall (Kühnert, 1993). Regelwerke, die Hierarchien von Assimilations- und Elisionsregeln aufstellen, können kombinatorische Effekte erfassen – wie etwa Schwa-Tilgungen, die erst zu Konsonantenverbindungen führen, die dann assimiliert oder elidiert werden können. Solche Regelwerke bilden aber nicht artikulatorisches Planungsverhalten ab. Da phonetische Informationen trotz Tilgung von Segmenten noch bestehen bleiben können (Kohler, 2003b), werden im empirischen Teil auch Vokalnasalierung und Laryngalisierung, die im Korpus notiert sind, untersucht.

Assimilationen, also Angleichungen von benachbarten Konsonanten in Artikulationsort, -art oder Stimmhaftigkeit, treten wie Elisionen nicht in allen Positionen auf. Angleichungen im Ort betreffen im Deutschen z. B. alveolare Konsonanten mit Ausnahme von Frikativen. Als mögliche Ursachen für stellungsbedingte Effekte führt Kohler Ökonomiegründe an: Apikale Zungenbewegungen seien sehr aufwändig und daher potentielle Kandidaten für Reduktionen und – ganz im Sinne der H&H Theorie (vgl. dazu Kapitel 3.1.1) – würden in perzeptiv wichtigen Positionen (Onset, Frikative in der Koda) nicht verändert. Es gibt weitere Fälle, in denen schwache Formen nicht auftreten. Kohler gibt das Beispiel von

ihr, das als Personalpronomen durchaus bis zum [ɐ] reduziert werden kann, während es sich als Possessivpronomen robust zeigt.

In der Regel wird das Auftreten von Elisionen und Assimilationen mit informellem Sprechstil, Funktionswörtern und hoher Wortfrequenz in Verbindung gebracht. Diese Bedingungen fallen dann häufig mit hohem lokalen Tempo bzw. geringeren Wortlängen zusammen. Analysen zeigen, dass in der Domäne der Silbe vor allem Kodakonsonanten und unbetonte Silbenkerne von solchen Veränderungen betroffen sind (Greenberg, 1999; Greenberg et al., 2003a). Für den Fall von Elisionen ist eine geringere Wortlänge sogar eine notwendige Folge, da nicht gleichzeitig von einer Längung der restlichen Segmente ausgegangen wird. Tatsächliche Analysen von tempobedingten Reduktionen auf Symbolebene sind selten. In einer der wenigen Ausnahmen (Fosler-Lussier und Morgan, 1999) zeigt sich, dass eine hohe Silbenrate auf Phonebene mit mehr Aussprachevarianten und stärkerer Abweichung dieser Varianten von einer kanonischen Aussprache korreliert: Bei der Analyse der 117 häufigsten Wörter (über 40 Tokens) zeigen alle Typen, die signifikante Tempounterschiede zwischen schnellen und langsamen Realisierungen aufweisen, auch eine signifikant geringere Anzahl kanonischer Aussprache für schnellere Token. Solche Abweichungen von einer kanonischen Aussprache betreffen unbetonte Silben stärker als betonte, sowie Silben mit Kodakonsonanten häufiger als solche ohne Koda. Jedoch zeichnen sich nicht alle Wörter durch diese Systematik aus. Etwa die Hälfte der untersuchten Silben und ein Drittel der Wörter zeigen tempobedingte Ausspracheabweichungen. Dabei weisen Wörter dann keinen Zusammenhang zwischen der Stärke ihrer Variation in den Realisierungen und ihrem Tempo auf, wenn ihre Auftretensfrequenz insgesamt geringer ist. Die Untersuchung der zehn häufigsten Funktionswörter aus der Studie von Bell et al. (2003) ergibt, dass tempobedingte Reduktionen und Elisionen von der Identität des Wortes abhängen: Während sich *„I, a, the, to“* stark mit ansteigendem Tempo verändert, weisen *„that, it, in“* keine Korrelation zwischen Tempo und Reduktionsphänomenen auf.

In einer Forschungsarbeit zur Optimierung eines Sprachsynthesesystems des Deutschen wird die genaue Aussprache von Wörtern in Abhängigkeit von über Kontextinformationen geschätzten Wortdauern variiert. Dadurch werden die ausgewählten Varianten zwar informeller, da sie eher Reduktionen und Elisionen beinhalten. Aber die Varianten werden von Probanden nur bedingt natürlicher oder verständlicher bewertet (Werner et al., 2005, 2003).

Trouvain et al. (2001) untersuchen den Einfluss der Artikulationsrate auf die Aus-

sprache im Deutschen. Bei erhöhter globaler Phonrate (für die Domäne zwischen zwei Pausen) kommt es zu mehr Elisionen und Reduktionen, wobei Elisionen mit der Ausnahme von /n/-Reduktionen weit häufiger auftreten.

Da Sprechgeschwindigkeit, Worthäufigkeit und Wortart korrelieren, wird im empirischen Teil keine allgemeine Wahrscheinlichkeit vom Abweichen einer kanonischen Aussprache bei höherem Tempo überprüft. Grund ist der geringe Erkenntnisgewinn über einen direkten Zusammenhang. Stattdessen werden häufige Wörter ausgewählt, da hier ausreichend Realisierungen eines Wortes vorhanden sind, um ihr Auftreten mit Tempovariation zu korrelieren. Insbesondere die Verwendung eines an der Silbenrate angelegten Maßes (vgl. Kapitel 6) soll bisherige Untersuchungen zum Deutschen erweitern, die sich nur auf Phonrate oder Wortlänge stützen.

4 Sprechtempo als Teil des Sprachverstehens

In den vorangegangenen Kapiteln wurden akustische Effekte von Sprechgeschwindigkeit auf Aussprachevariationen dargestellt. Die Unterschiede in der Aussprache führen teilweise zu unterschiedlichen Urteilen oder Leistungen bei Identifikationsexperimenten. Schnell artikulierte Sprache kann allerdings genauso verständlich sein wie solche, die mit normalem Tempo produziert wurde (vgl. z. B. Krause und Braida, 1995). Für normale Tempovariation eines Sprechstils ergeben sich keine Unterschiede in der Verständlichkeit (Bradlow et al., 1996). Allerdings zeigen automatische Spracherkenner starke Leistungseinbußen bei tempobedingten Variationen, wenn das Sprechtempo nicht berücksichtigt wird (Chung und Seneff, 1999). Menschliche Spracherkennung ist prinzipiell nicht dieser Beeinträchtigung unterworfen, wie sowohl die in Kapitel 2.3 gezeigte Adaptionfähigkeit, also auch die Nutzung von Kontextinformationen (vgl. Bard et al., 2001) zeigen. Dieser robuste Umgang mit Variation motivierte zu zahlreichen Untersuchungen zur Thematik großer akustischer Variabilität bei gleichzeitig invarianten (Wort-)Bedeutungen auf kognitiver Ebene.

In analytischen Modellen zur Sprachverarbeitung wird dieses Problem theoretisch gelöst, indem davon ausgegangen wird, dass akustische Informationen in einem frühen Stadium der Verarbeitung in einer komplexen Weise linguistischen Einheiten, wie etwa Phonemen, zugeordnet werden. Aus diesen Einheiten wird wiederum unter Einbeziehung syntaktischer und semantischer Regeln in einem Lexikon eine (Wort-)Bedeutung ermittelt. Akustische Variabilität wird in solchen Systemen als unerwünschte und unnütze Abweichung durch den Artikulationsprozess angesehen (vgl. auch Chomsky und Halle, 1968). Bei der kognitiven Weiterverarbeitung gehen kontextuelle Informationen verloren, da sie für eine Spracherkennung weder notwendig sind, noch die Kapazitäten ausreichen, um diese zu speichern (Posner, 1964; Joos, 1948). So ist es durch die zugrunde liegende Vorstellung notwendig, von einem Normalisierungsprozess auszugehen (Pisoni, 1997). Drei bedeutende Ursachen von akustischer Variabilität sind Sprecher, linguistischer Kontext und Sprechtempo. Letzteres wird zumindest seit-

dem dazugezählt, seit deutlich wurde, dass Tempo nicht nur sprecherspezifisch ist, sondern auch intrapersonell sehr stark variiert (siehe Kapitel 6), was auch im empirischen Teil deutlich wird (siehe Kapitel 8).

Auf dieser theoretischen Basis wurde versucht, invariante Parameter zu finden, die Phonemidentität charakterisieren. Akustische Parameter streuen aber so stark, das sich nicht immer eindeutig Phone klassifizieren lassen, umso mehr, wenn Kontext, prosodische Position und Tempo variieren. Deshalb werden Informationen über Tempo in intrinsische Parameter mit eingebunden (Perkell und Klatt, 1986). Diese Sichtweise wird im folgenden Kapitel in Frage gestellt. Die daraus motivierten Experimente liefern dennoch wichtige Erkenntnisse über die Relevanz von Informationen über die Sprechgeschwindigkeit bei der menschlichen Sprachverarbeitung.

4.1 Zur „intrinsisch"–„extrinsisch" Unterscheidung

In seinem eher biographisch ausgerichteten historischen Überblick über phonetische und phonologische Forschung spezifiziert Fant (2004) den ungefähren Zeitraum von 1965 bis 1980 als solchen, in dem die Suche nach invarianten akustischen Parametern einen besonderen Themenschwerpunkt einnimmt. Ein Beispiel für die These, dass sich Sprachperzeption auf invariante Parameter der akustischen Repräsentation von Sprache stützen würde, bietet der Artikel von Stevens und Blumstein (1978). Inzwischen hat sich allerdings die Ansicht durchgesetzt, dass keine absolut invarianten akustischen Parameter existieren (vgl. dazu auch Assmann et al., 1982; Perkell und Klatt, 1986). Zu verschieden sind untersuchte Parameter in unterschiedlichen Bedingungen und zu groß ist die Variabilität der Parameterwerte. Statt dessen gerät zunehmend akustische und auch artikulatorische Variabilität und die Systematik in ihrem Auftreten in den Fokus von Untersuchungen. Diese Variabilität zeichnet sich dadurch aus, dass sie selbst bedeutsame Informationen trägt damit gesprochene Sprache natürlicher klingen und einfacher kognitiv verarbeiten lässt. Hawkins und Smith (2001) nehmen in diesem Rahmen speziell Bezug auf das Sprechtempo:

> Such simple representations of phonological form may be encouraged by a further tendency to assume that lexical and even sentence meaning remains unchanged no matter how fast the speech: invariance of form and meaning if not of the physical signal. In reality, an important aspect of meaning in normal conversation includes so-called paralinguistic information like overall rate and

> rate changes, but even if we ignore this, acoustic-phonetic models of speech perception must be able to account for variations in rate of speech, if only because of the segmental reorganisation that typically accompanies them. Time as rate is therefore crucial. (S. 18)

Anstatt akustische Sprechraten zu verwenden, wird von Miller und Liberman (1979) die Geschwindigkeit von artikulatorischen Gesten als mögliche Lösung nicht-existierender invarianter akustischer Parameter vorgestellt. Ähnlich argumentiert Fowler (1980). Sie kritisiert sogenannte Sprachproduktionsmodelle mit extrinsischem *Timing*. Damit meint sie Modelle, die einen artikulatorischen Plan annehmen, der von zeitlosen, seriell angeordneten Planungseinheiten wie Phonemen oder distinktiven Merkmalen ausgeht, bei denen Dauern für eine Definition von Planungseinheiten keine Rolle spielen. Damit ist eine noch nicht artikulierte mentale phonologische Form gemeint. Für solche Modelle sind zeitliche Informationen zu den Segmenten unerheblich, da von diskreten Einheiten ausgegangen wird. Solche Modelle könnten aber Effekte von Koartikulation oder Abhängigkeit der *VOT* von umgebenden Dauern nur beschränkt erklären, da sie zwar artikulatorische Einflüsse über segmentalen Kontext und mit Hilfe unabhängiger distinktiver Merkmale zulassen, aber keine echte Modellierung von Koartikulation (als ko-produzierte Gesten) vornehmen.

Diese systematische Variation wird von extrinsischen Modellen nicht erfasst. Fowler argumentiert, dass ein Produktionsmodell *Timing*-Informationen als integralen Teil der Produktionseinheiten modellieren muss, um den empirisch gefundenen Daten gerecht zu werden. Der zeitliche Verlauf der Sprachproduktion wäre demnach in einer Definition artikulatorischer Gesten enthalten (intrinsisch). Weil Menschen mit akustischen Informationen und eigenem artikulatorischen Wissen auf die Produktion rückschließen könnten, würden *Timing*-Informationen, speziell Dauerverhältnisse innerhalb von Silben, auch bei der Perzeption benutzt (Fowler, 1986).

Kritik an dieser Theorie gründet sich auf Experimente, in denen artikulatorisches Wissen bei der Sprachverarbeitung ausgeschlossen werden kann, da nichtsprachliche Stimuli zu vergleichbaren Ergebnissen führen (Pisoni et al., 1982). Für die Vokalartikulation ergeben sich auch gravierende individuelle Unterschiede, die invariante Planungseinheiten fraglich werden lassen (Johnson et al., 1993). Ob Sprache nun einen besonderen Verarbeitungsmechanismus aufweist oder nicht, ist eine Frage, die bereits vielfach diskutiert wurde (siehe etwa Liberman et al., 1967; Fowler, 1990). Die Verarbeitung von invarianten Gesten*definitionen*

anstatt ihren Ausführungen stellt eine Möglichkeit dar, auf kognitiver Ebene mit artikulatorischer und akustischer Variabilität zurechtzukommen (Johnson, 1997).

Hier bedarf es einer Anmerkung zur extrinsisch/intrinsisch Unterscheidung. Die gerade vorgestellte Definition intrinsischer zeitlicher Informationen bezieht sich auf den Bereich der Sprachproduktion, die durchaus auf die Wahrnehmung übertragen werden kann (Fowler, 1986). Es stellt sich aber die Frage, in welcher Form diese Tempoinformationen verarbeitet werden: Intrinsisch, also direkt bei der Verarbeitung lokaler akustischer Eigenschaften; oder extrinsisch und damit auch zeitlich später. Pompino-Marschall et al. (1982) konnten zeigen, dass die Silbenrate ein wichtiger Parameter für die Wahrnehmung lokalen Tempos darstellt. Nicht die Dauer der geschlossenen Silbe, sondern die Abstände sogenannter P-Center (Pompino-Marschall, 1990) bilden dabei relevante Informationen für die Tempoverarbeitung. Diese P-Center liegen zeitlich in dem Bereich, wo prä-vokalischer Konsonant und Vokal artikulatorisch ko-produziert werden. Da das im empirischen Teil verwendete Tempomaß perzeptiv bestätigt ist und die Silbenrate stark mit in dieses Maß eingeht, entspricht es ausreichend wahrgenommener Sprechgeschwindigkeit. Dieses Tempomaß steht durchaus mit einer Theorie intrinsischen *Timings* im Einklang, auch wenn die akustische Verarbeitung von Tempo dann ja nicht intrinsisch (z. B. über die Steilheit von Formanttransitionen), sondern extrinsisch als Abgleich von Tempo als Silbenrate oder Geschwindigkeit von Artikulationsbewegungen mit spektralen Eigenschaften durchgeführt wird.

4.2 Zeitliche Domänen der Tempoverarbeitung

Es wird deutlich, dass bei der Sprachverarbeitung Informationen über das Tempo mit in die Identifikation linguistischer Einheiten eingehen müssen. Hierbei scheinen sowohl lokale als auch globale Informationen auf unterschiedliche Art und Weise Verwendung zu finden (vgl. Kapitel 2.3). Ergebnisse von Perzeptionsexperimenten suggerieren eine parallele Verarbeitung mit zwei Zeitfenstern. Eines liegt im Bereich von etwa 30 ms und kann daher dem phonetischen Segment[1] zugeordnet werden, die andere bei mindestens 300 ms, was der Silbe entsprechen könnte (Chait et al., 2005). Diese Zeitfenster werden in einem gewissen Rahmen als variabel angesehen.

Die Annahme von solchen Zeitfenstern, mit denen akustische Informationen er-

[1] Oder akustischen Korrelaten der Artikulation von Einheiten, die phonologisch mit dem Phonem erfasst werden.

fasst werden, um dann später einer Bedeutung zugeordnet zu werden, kann z. B. erklären, dass Sprachsignale, die in bis zu 50 ms Intervallen umgekehrt dargeboten werden, dennoch verstanden werden (Saberi und Perrott, 1999). Solche Zeitfenster anzunehmen erklärt auch stellungsbedingte Effekte wie den Einfluss nachfolgender Informationen auf Segmentidentifikation wie etwa Phonotaktik (Massaro und Cohen, 1983b), da die tatsächliche Interpretation der akustischen Daten erst später stattfindet. Auch Newman und Sawusch (1996) interpretieren ihr Ergebnis, dass Segmentdauern nachfolgender, aber nicht direkt benachbarter Konsonanten, Identifikationen beeinflussen können (z. B. die Klassifizierung eines Stimulus als /tʃ/ gegenüber /ʃ/) als Zeichen für ein zeitliches Fenster lokaler Tempointegration mit maximal 300 ms Abstand zum betroffenen Phon, was bei Annahmen von Symmetrie wiederum einem Maximum von etwa 600 ms entspricht. Der von ihnen beschriebene Einfluss ist unabhängig von Phonotaktik oder akustischer Ähnlichkeit beider Segmente. Dass Summerfield (1981) keinen Wahrnehmungswechsel von Zielkonsonanten durch Dauermanipulation eines silbenfinalen Frikativs hervorrufen konnte, erklären sich Newman und Sawusch mit einem über diese Fensterlänge hinausgehenden zeitlichen Abstand.

Da bei der Annahme von zwei verschiedenen Zeitfenstern segmentale und prosodische Informationen erst einmal getrennt von einander extrahiert werden, erübrigt sich die Frage nach der wichtigsten Einheit bei der Sprachperzeption, da es mehrere gleichwertige geben kann. Extrinsische Verarbeitung von Tempoinformationen, etwa in Form von Adaption oder Normalisierung, lassen sich über einen späteren Abgleich der beiden parallelen Informationsströme mit segmentalen und prosodischen Informationen, die beide Fensterlängen mit sich bringen, erklären.

Nach dieser Sichtweise wäre der Einfluss der globalen Rate, wie er in vielen Experimenten nachgewiesen wurde, kritisierbar. Ein Beispiel dazu ist die Fußnote 3 in Miller (1981), in der die lokale Rate zwar als bedeutsam bezeichnet wird, in ihrem Review aber nicht berücksichtigt wird. Dazu sei an dieser Stelle angemerkt, dass die zitierten Untersuchungen alle Tempoauswirkungen auf Phonebene zum Thema haben. Da die Stimuli weitgehend über Trägersätze eingesprochen wurden (mit qualitativen Instruktionen zum Tempo), ist es präzise, diese Tempounterschiede als global zu bezeichnen. Die Wirkung kann jedoch auch auf eine lokale Sprechgeschwindigkeit zurückgeführt werden, da die Stimuli selbst meist sehr kurz sind und sich die Tempi direkt auf die Silbendauern der Zielsilbe auswirken. Selbst benachbarte Silbendauern werden so zu globa-

lem Tempo gezählt. Eine der wenigen Ausnahmen bildet Wayland et al. (1994), die zeigen, dass unter gleichen lokalen Bedingungen der weitere Kontext, also die globale Rate, auch einen Einfluss auf die Wahrnehmung hat. Sie diskutieren, ob es sich nicht um zwei verschiedene Mechanismen handelt, lokale Rate, die Indiz für ein intrinsisches *Timing* ist, und das Satztempo. Effekte, die globalem Tempo zugeordnet werden, könnten durch das größere Zeitfenster erklärt werden, da kurze einsilbige Wörter als Stimuli verwendet werden, die weniger als 300 ms lang sind. Ähnliches gilt auch für Kidd (1989), der einen Einfluss von Tempo vorangehender Silben sogar über eine Phrasengrenze hinweg auf die Wahrnehmung von Stimmhaftigkeit silbeninitialer velarer Plosive nachweist. Als Effekte globaler Sprechgeschwindigkeit mit wirklich großen Domänen wie etwa Sätzen können diese Ergebnisse allerdings nicht angesehen werden.

Konkrete Modelle des Prozesses der Einbeziehung von Sprechtempoinformationen bei der Identifikation von Phonemen existieren kaum. Eine Ausnahme ist *PHONET* (Boardman et al., 1999), das im Rahmen der *Adaptive Resonance Theory (ART)* (vgl. Grossberg, 2003) entwickelt wurde. Dieses neuronale Modell kann die Identifikationsleistungen von Konsonant-Vokal Transitionen für das /ba/–/wa/ Kontinuum (Miller und Liberman, 1979; Schwab et al., 1981) fast komplett simulieren. Dabei interagiert die (bottom-up) Verarbeitung von neuronalen Signalen, die von akustischen Reizen ausgelöst werden, mit einer (top-down) Verarbeitung von Erwartungen, die gelernten Kategorien entsprechen. Im Modell führt diese Interaktion zu neuronalen Resonanzen, deren Verarbeitungsgeschwindigkeit durch Informationen über das Sprechtempo, z. B. Transitionsbewegungen, gelenkt wird, um zu einer tempounabhängigen Repräsentation phonemischer Einheiten zu kommen. Dafür werden zwei parallele Kanäle angenommen, die verschiedene akustische Informationen filtern. In einem Kanal werden relativ stationäre Informationen, wie *steady states* von Vokalen, verarbeitet, in dem anderen Kanal nicht-stationäre, wie Transitionen. Im Arbeitsgedächtnis stehen die so gefilterten Informationen in asymmetrischer Beziehung zu einander. Tempovariation aus den transienten Informationen beeinflusst die neuronale Aktivierungsstärke und damit die Integrationsgeschwindigkeit auch für den Bereich für stationäre Informationen. Insgesamt werden dadurch akustische Parameter relativ zur Sprechgeschwindigkeit verarbeitet. Mit diesem komplexen Modell wird eine recht späte Klassifikation und Identifikation vorgenommen, die auch Effekte nachfolgender Vokaldauern auf Klassenzugehörigkeiten ermöglicht, wie sie u. a. von Miller und Liberman (1979) berichtet werden. Soweit dem Autor bekannt, existieren abgesehen von ART jedoch keine Implementierungen

für allgemeinere Effekte von Tempoverarbeitung, sehr wohl dagegen Ansätze zur Modellierung sprecher- und kontextspezifischer Verarbeitung (siehe unten).

Da Dauern von Segmenten relativ zu ihren Nachbarn wahrgenommen werden – so wirkt etwa ein Laut kürzer vor längeren Nachbarn –, könnte dies für temporale Informationen eine Art direkte und unbewusste Temponormalisierung darstellen (Miller und Liberman, 1979). Der Prozess wird von Miller und Dexter (1988) und Sawusch und Newman (2000) als früh und relativ unbewusst eingeordnet. Andere Experimente geben allerdings Hinweise darauf, dass er nach einer Segmentation des akustischen Signals, aber vor einer Identifikation der Segmente stattfindet (Green et al., 1994).

In einem neueren Experiment zeigt sich, dass ein langer Vokal durchaus bei hohem Tempo als zwei Segmente interpretiert wird. Abhängig von dieser Segmentation ergeben sich verschiedene Segmentdauern, die relativ zur Umgebung wiederum Einfluss auf Phonemidentifikationen haben (Brenner-Alsop, 2006). Auch plötzliche F_0- oder Formantveränderungen beeinflussen die Segmentation und damit auch die zur Normierung herangezogenen Segmentdauern mitsamt ihrem Einfluss auf die Identifikation anderer Segmente (Newman und Sawusch, 1996).

Insgesamt mehren sich die Hinweise darauf, dass Tempoverarbeitung – ähnlich wie Sprechererkennung mit ihrem Einfluss auf phonetische Klassifikationen – früh und unbewusst stattfindet, nämlich zeitlich vor einer linguistischen Identifizierung. Es wird davon ausgegangen, dass es sich dabei um einen aktiven Prozess handelt (vgl. Nusbaum und Magnuson, 1997; Mullennix, 1997). Hinweise für diese Interpretation kommen von Worterkennungsexperimenten, bei denen erhöhte Reaktionszeiten und schlechtere Leistungen durch wechselnde Sprecher und variierendes Tempo induziert wurden. Die Verarbeitung solcher Variationen reduziere kognitive Leistungen und könne damit kein passiver, automatischer Prozess sein, sehr wohl aber ein obligatorischer.

Im Zuge einer Theorie episodischen Lexikons (vgl. Goldinger, 1998), ließen sich diese Effekte ohne akustisch invariante Parameter erklären. Hier wird das Lexikon nicht minimalistisch, sondern als Sammlung einzelner Instanzen oder Exemplare mitsamt ihren nicht-linguistischen Auftretensinformationen angesehen. Ein Indiz dafür sind Erinnerungen an Einzelheiten beim Hören und Lesen von Wörtern, wie etwa dem Platz auf der Seite oder Geschlecht des Sprechers. Der Zugriff auf eine Bedeutung geschehe dabei über Analogie, sodass statt einer

komplexen Verarbeitung akustischer Informationen zu Phonemen (vgl. Kapitel 4), ein Vergleich mit ähnlichen gespeicherten Instanzen vollzogen würde. In einem solchen Modell kann nicht von Normalisierung von Sprecher- oder Tempovariation gesprochen werden. Stattdessen würden gerade diese (para- und extralinguistischen) Variabilitäten helfen, als „Indexinformationen" die passenden Bündel von Exemplaren zum Vergleich heranzuziehen. So sind beispielsweise bekannte Stimmen, oder ihnen akustisch ähnliche, einfacher zu verstehen als fremde. Dass feine sub-phonetische Informationen wie etwa genaue Dauern, oder prosodische, wie Nasalierung umgebender Segmente, bei der Wahrnehmung bedeutsam und kein „Rauschen" darstellen, beschreiben u. a. Hawkins und Smith (2001).

Zu sprecherspezifischen Unterschieden schreiben Nusbaum und Magnuson (1997):

> If talker normalization is needed to address a nondeterministic mapping between acoustic properties and linguistic categories, it cannot operate as a passive filtering mechanism, as implied by Palmeri et al. and Nygaard et al. Instead, it must actively test hypotheses about talker's vocal characteristics. These could be derived from context or from the utterance itself (Ainsworth, 1975). But this information does not need to modify the auditory representation of an utterance as a precursor to recognition of that information. (S. 124)

Diese Ansicht gilt auch für Tempovariation, die im gleichen Aufsatz häufig mit Sprechervariation verglichen aber als wenig beachtet bezeichnet wird (S. 20–121). Wenn Informationen über die Sprechgeschwindigkeit vor der eigentlichen linguistischen Verarbeitung oder Bedeutungszuweisung aus dem Sprachsignal extrahiert werden, muss dies allerdings nicht separate Prozesse bedeuten (vgl. Mullennix, 1997). Johnson (1997) entwirft ein episodisches Modell für Vokale und bekräftigt, das auch Sprechtempo in einem solchen Modell einen bedeutenden „Indexfaktor" darstellt:

> Variation in the speech signal caused by changes in speaking rate would be handled in the same way (including vowel reduction and even resyllabification and extensive gestural reorganization). So, although I have focused on talker variability in this chapter, I am aiming for a general model that uses the same mechanism to handle many different sources of variability in the speech signal. (S. 162)

Ein solcher Vergleich von Tempo und individuellen Unterschieden als „Index-

informationen“ würde eine extrinsische Verarbeitung von Tempo bedeuten. Innerhalb dieser Theorie ist aber auch eine intrinsische Verarbeitung denkbar, wenn Tempo zusammen mit spektralen und temporalen Charakteristika in einem vorausgewählten Bündel von Exemplaren ein Kriterium für den Analogietest darstellt. In Anbetracht der Ähnlichkeit im Einfluss von Tempo- und Sprechervariation auf Wortwahrnehmung erklärt sich jedoch die Gleichstellung beider Faktoren als „Indexinformation“.

Unabhängig vom Perzeptionsmodell ist deutlich geworden, dass Variation in der Sprechgeschwindigkeit nicht nur mit der auf akustischer und artikulatorischer Ebene zusammenfällt, sondern auch bei der Verarbeitung genutzt wird. Unter diesem Aspekt kann jeder im empirischen Teil ermittelte signifikante tempobedingte Unterschied eine Aussprachevariation darstellen, zu deren Verarbeitung auch Tempoinformationen genutzt werden müssen.

Teil II

Empirische Untersuchung

5 Zielsetzung und Durchführung

Ausgehend von den Darstellungen im theoretischen Teil ergibt sich die generelle Frage, ob und wie sich bisherige Ergebnisse tempobedingter Aussprachevariationen im Deutschen und speziell in deutscher Spontansprache wiederfinden. Es wurden bislang vorrangig Untersuchungen durchgeführt, in denen das Sprachmaterial aus gelesenen Texten oder aus einzelnen Sätzen bestand. Sprechgeschwindigkeit wurde zumeist auf globaler Ebene variiert, entweder per Instruktion der Versuchsleiter, oder indem Vorleser nachträglich in schnelle und langsame Sprecher eingeteilt wurden.

In diesem empirischen Teil der vorliegenden Arbeit werden Durchführung und Ergebnisse akustischer Analysen deutscher Spontansprache (vgl. Kapitel 7) dargestellt. Fokus dieser Analysen liegt auf Zusammenhängen von *lokaler* Sprechgeschwindigkeit (siehe Kapitel 6) und Aussprache. Damit soll insbesondere die Frage beantwortet werden, ob Sprecher systematische Veränderungen in ihrer Aussprache aufweisen, wenn sie ihr Tempo verändern. Unterschiede zwischen global langsamen und schnellen Sprechern werden hier nur in dem Fall von Vokalen zusätzlich untersucht (vgl. Kapitel 3.1.1), da sie bedeutsam für die Einschätzung von Variation innerhalb der Sprecher sind. Ansonsten wird Variation in der Sprechgeschwindigkeit relativ zu jedem Sprecher behandelt, da diese intrapersonelle Tempovariation im Korpus weit größer ist als interpersonelle, sodass die Unterschiede zwischen langsamen und schnellen Sprechern im Vergleich dazu kaum ins Gewicht fallen (vgl. Kapitel 8.1).

Dieser empirische Teil umfasst zum einen spektrale Analysen von Phonen: Temporale Parameter werden hierbei nicht betrachtet, da diese bereits ausgiebig in anderen Studien untersucht wurden. Um spektrale Eigenschaften von Phonemrealisierungen zu erfassen, werden solche Phone berücksichtigt, die in der Annotation sowohl auf kanonischer Ebene wie auch in der tatsächlichen Aussprache mit dem gleichen Symbol annotiert sind. Grundlegende Hypothese bildet die Annahme, dass sich mit steigendem Tempo signifikante Veränderungen ergeben, die als Ausdruck artikulatorischer Reduktion gewertet werden können. Diese Analysen werden separat für verschiedene Bedingungen durchgeführt, um

bereits bekannte Interaktionen von Tempo und Reduktion mit Betonung und Wortart auszuschließen, und stattdessen Aussprache auf Effekte *innerhalb* solcher Bedingungen zu überprüfen. Bei den Vokalen werden Diphthonge von der Untersuchung ausgeschlossen, da hier bereits ein aussagekräftiges Ergebnis vorliegt (vgl. Kapitel 3.1.2); stimmlose Frikative sind ausgewählt worden, weil mit dem *COG* ein sinnvoller und robust zu messender Parameter als Korrelat von Reduktion verwendet werden kann (vgl. Kapitel 3.1.3). Bei den Monophthongen steht im Vordergrund, nach einem uneinheitlichen Bild in der Literatur mit sich widersprechenden Ergebnissen, zu überprüfen, welches beobachtete Resultat auf die hier untersuchte Sprache und den untersuchten Sprechstil zutrifft. Dem gegenüber ist dem Autor bei den Frikativen überhaupt nur eine Untersuchung zu dieser Fragestellung bekannt, sodass es hier um eine Verbreiterung der empirischen Basis geht.

Zum anderen werden Wortrealisierungen anhand der symbolischen Umschrift untersucht. Auch hier ist der Gegenstandsbereich die phonetische Realisierung. Jedoch stehen in diesem Fall segmentelle Prozesse im Vordergrund, also Abweichungen von einer kanonischen Form, die mit einem neuen Symbol annotiert sind. Obwohl die Domäne das Phonem bleibt, wird diese Analyse im Rahmen von Wortrealisierungen durchgeführt, um damit den phonetischen Kontext zu kontrollieren, was für diese Daten zwingend notwendig ist.

Nach der Darstellung des verwendeten Korrelats für lokales Sprechtempo (Kapitel 6) wird das Korpus beschrieben, das für die Untersuchung herangezogen wurde (Kapitel 7). Danach folgen die einzelnen Analysen, wobei die spezifische Fragestellung jeweils im betreffenden Kapitel angegeben wird. Letzteres gilt auch für die dafür verwendeten statistischen Methoden. Diese müssen den jeweiligen Datensätzen und Fragestellungen angepasst werden, da die Daten nicht für die Analyse in einem speziellen Experiment erhoben wurden, sondern es sich aus statistischer Sicht um Felddaten handelt. Damit gilt auch der Interpretation signifikanter Effekte besondere Aufmerksamkeit. Die teilweise hohen Fallzahlen führen zu niedrigen p-Werten, weswegen das α-Niveau entsprechend angepasst werden muss. Ein hier häufig verwendetes Verfahren betrifft gerade diese größeren Datensätze in Kapitel 8 und 9 mit ihren Fallzahlen im 3-stelligen Bereich innerhalb einer Bedingung. Diese Daten sind durch fehlende Werte in einigen Bedingungen, verschiedene, teilweise nicht zu kontrollierende Kontexte und bei fehlender Berücksichtigung mancher Kontexte (z. B. konsonantische Umgebung von Vokalen) auch unterschiedliche Anzahlen von „Wiederholungen“ in

einer Bedingung gekennzeichnet. Hier ist die Homogenität der Varianzen bei den Daten vorhanden und ihre Ähnlichkeit zur Normalverteilung lässt parametrische Tests zu. Da aber die Gruppengrößen (u. a. zwischen einzelnen Phonemen) stark schwanken und Mittelwertbildungen innerhalb dieser Gruppen die Einschätzung von Varianzen ausschließen, werden weitgehend statt klassischer linearer Modelle oder Varianzanalysen *lineare gemischte Modelle* (Pinheiro und Bates, 2001) mit dem *R*-Paket[1] *nlme* berechnet. Diese reagieren robust auf Verletzungen der Voraussetzung der Normalverteilung und nicht ausbalancierte Modelle und ermöglichen die Modellierung von Zufallsfaktoren (wie Personen), multi-level-modelling (Venables und Ripley, 2002) und (gegenüber nicht-parametrischen Tests) Parameterabschätzungen. Wird kein Zufallsfaktor oder Variablenhierarchie angegeben, handelt es sich bei Tests auf die F-Verteilung um klassische Varianzanalysen, ansonsten um eine Statistik mit *nlme*. Falls die Voraussetzungen für diese Statistiken nicht gegeben sind, werden entsprechende nichtparametrische Verfahren verwendet und im Text benannt. Die Darstellung von Ergebnissen statistischer Tests erfolgt nach den Richtlinien der *American Psychological Association*.

[1] `www.r-project.org`

6 Maße für das Sprechtempo

Wie in den vorangegangenen Abschnitten beschrieben, gibt es viele Möglichkeiten, Sprechgeschwindigkeit zu definieren. Eine qualitative Möglichkeit, die oft für Laborsprache verwendet wird, besteht darin, normale, schnelle und langsame Tempi zu unterscheiden. Normale Sprechgeschwindigkeit zeichnet sich daher entweder durch durchschnittliche Werte aus oder ist das Resultat einer expliziten Sprechanweisung *„normales Tempo“*. Problematisch ist dies insofern, dass die Klassengrenzen je nach verwendeten Sprachaufnahmen und Entscheidung der Forscher schwanken können, da die genauen Grenzen zwischen *„normal“* und etwa *„schnell“* individuell festgelegt werden.

Da Segmentlängen aufgrund der großen Kontextvariabilität unpraktikabel sind, hat sich ein klassisches Maß etabliert: Das *globale Sprechtempo*, das als Silben- oder Phonrate über die aufgenommenen Domänen (Äußerung, Phrase oder Satz) gemittelt wird. Dabei muss zwischen der allgemeinen Sprechgeschwindigkeit, gemessen als Segmente über Äußerungsdauer, und der Artikulationsrate, die die Pausen ausschließt, unterschieden werden, denn das Pausenverhalten ist selbst höchst variabel (Goldman-Eisler, 1968; Kowal, 1991). In der vorliegenden Arbeit wird grundsätzlich die Artikulationsrate betrachtet, da lokales Tempo bei Stille keinen Sinn ergibt und gefüllte Pausen nicht unter die Aussprachevariationen fallen. Das bedeutet natürlich nicht, dass Pausen irrelevant wären. Ihre Verteilung gibt zum Beispiel Hinweise darauf, dass es bei erhöhtem Tempo zu einer Umstrukturierung zugunsten weniger prosodischer Einheiten kommt (Trouvain und Grice, 1999; Shattuck-Hufnagel und Turk, 1996), und ihr Auftreten kann zu einer anderen Wahrnehmung von globaler Sprechgeschwindigkeit führen (Lass, 1970). Wenn also von Sprechtempo oder Geschwindigkeit gesprochen wird, ist die Artikulationsrate gemeint. Dass sich diese sprachliche Präzision nicht durchgesetzt hat, sondern allerlei Synonyme für Artikulationsrate verwendet werden, verdeutlicht ein Blick in das Literaturverzeichnis. Ein Grund dafür mag sein, dass die Artikulationsrate so dominant verwendet wird, dass in vielen Fällen – wie auch dem vorliegenden – eine Differenzierung zur Sprechrate unnötig erscheint.

Das Fehlen präziser Definitionen und deren konsistenter Verwendung für Maße von Sprechgeschwindigkeit zeigt, dass dessen lokaler Charakter im Großteil der wissenschaftlichen Forschung nicht berücksichtigt wurde. So sind globale Maße weit verbreitet, da die Artikulationsrate als individuell und wenig variabel angesehen wurde (Goldman-Eisler, 1968).

Das durchschnittliche Tempo im Deutschen beträgt etwa $5-8\,^{\text{Silben}}/_{\text{s}}$, oder auch $10-15\,^{\text{Laute}}/_{\text{s}}$ (Pompino-Marschall, 1995), wobei Sprachen mit weniger komplexen Silbenstruktur andere Tempi aufweisen können. Die enorme Spannbreite der Werte erklärt sich u. a. durch die bereits aufgeführten Einflüsse auf das Tempo.

Aufgrund der Unterschiede in der Silbenkomplexität im Deutschen erklärt Essen (1949) die Phonrate als adäquateres Maß als die Silbenrate. Dem gegenüber wird gerade für längere Abschnitte die Silbenrate gern als Korrelat für globales Tempo verwendet. Dies mag seine Ursache darin haben, dass die Silbe als eine der primären Domänen der Artikulation angesehen wird. Beispielsweise organisieren sich Segmentdauern stark innerhalb von Silben. Begründungen für die Verwendung des einen oder anderen Maßes sind nicht verbreitet. Dass die Unterscheidung von Phon- und Silbenrate durchaus sinnvoll ist, zeigt die im lokalen Bereich nur mittelstarke Korrelation der beiden Maße miteinander (vgl. Pfitzinger, 1998, $r \approx .60$).

Als geeignetes Korrelat für die vorliegende Untersuchung wird die *perzeptive lokale Sprechrate* (PLSR) Pfitzinger (1999) verwendet. In seiner Arbeit zeigt Pfitzinger, dass die PLSR sehr stark mit der Wahrnehmung von Tempo im Deutschen korreliert. Bei diesem Maß handelt es sich um eine lineare Kombination aus Silben- und Phonrate, da beide auf die menschliche Einschätzung von Tempo einwirken. Dabei wird die Silbenrate leicht stärker gewichtet (etwa 55%). Die Segmentdauern beider Ebenen werden reziprok in ein nicht-stetiges Signal umgewandelt (im vorliegenden Fall 1 kHz Samplerate). Die hohe Datenrate hat zur Folge, dass die PLSR bei Pausen nicht zusätzlich auf Null gesetzt werden muss. Eine Glättung in stetige Signale erfolgte mit einer Hannfensterung (620 ms Breite für jedes Sample). Die Kombination der beiden Raten erfolgt nach Modell 1 ($r = .89$) aus Pfitzinger (2001), das auch in der vorliegenden Arbeit verwendet wird:

$$PLSR = 8,14 \cdot Silbenrate + 3,31 \cdot Phonrate + 6,07$$

Die sich langsam und stetig verändernde PLSR berücksichtigt auch die Silben-

komplexität und die lokale Charakteristik von Tempo. Durch eine Mittelung der Segmentraten in einem Zeitfenster von 620 ms werden starke Einflüsse des direkten Kontextes entfernt, ohne zu einem globalen Maß zu werden. Abgesehen von der perzeptiven Relevanz hat PLSR noch praktische Vorteile gegenüber anderen Maßen. Es ist direkt aus annotierten Korpora extrahierbar und ermöglicht die Zuweisung sinnvoller Werte zu jeder Art von zu untersuchenden Einheiten.

Es gibt noch weitere Einflüsse auf die Wahrnehmung von Sprechgeschwindigkeit, die nicht in die PLSR eingehen. Dazu gehören Pausen (siehe oben), die besonders für das globale Tempo wichtig sind, und die Grundfrequenz (Kohler, 1986), die in Pfitzinger (2002) zu einer leicht verbesserten Vorhersage führt (von $r = .906$ auf $r = .920$). Da die Evaluierung der PLSR prosodischen Kontext vernachlässigt, könnte der leichte Einfluss von F_0 sogar noch größer sein, falls Probanden die Grundfrequenz nicht in ihrer absoluten Ausprägung, sondern z. B. in Bezug auf die Intonationskontur verarbeiten. Bei Phrasen mit ähnlicher globaler Sprechgeschwindigkeit in geäußerten Phonen, führen verschiedene Raten von Phonemen, also einschließlich Elisionen, zu Unterschieden in der Tempowahrnehmung (Koreman, 2005). Dennoch stellt die PLSR ein ausreichendes Maß für Sprechtempo dar, das durch seine genannten Eigenschaften praktikabler ist als globale Raten und Lautdauern.

Während sich Sprechgeschwindigkeit durch seine Wirkung auf Dauern in allen prosodischen Ebenen widerspiegelt, lässt sich die PLSR als bewusst wahrgenommenes Tempo nach der Einteilung prosodischer Ereignisse in A-, B- und C-Prosodie (Tillmann und Mansell, 1980) der A-Prosodie zuordnen. Die Domäne der PLSR konstituiert sich aus durchgeführten Perzeptionstest auf über 600 ms, was dem Minimum der A-Prosodie nahe kommt. Dieses Maß ähnelt auch dem Ergebnis von Pickett und Pollack (1963), dass unter einer Dauer von 600 – 800 ms die Worterkennung von Stimuli rapide absinkt. Dabei ist nicht das Tempo wichtig, sondern alleine die Gesamtdauer eines Stimulus, da die mit erhöhtem Tempo einher gehende Reduktion durch mehr phonetischen Kontext ausgeglichen wird. Kato et al. (1997) zeigen, dass sich Tempowahrnehmung erst ab einer Dauer konstituiert, die bei mindestens eine Silbe liegen muss. Damit entspricht die in der PLSR verwendete zeitliche Domäne dem, was in Kapitel 4.2 bereits als ein Zeitfenster bei der Sprachverarbeitung dargestellt wurde.

7 Aufbereitung der Daten

Untersucht wird das Kiel Korpus spontaner Sprache (IPDS, 1995–1997), das im Rahmen des *Verbmobil*-Projekts (vgl. Wahlster, 2000) entstanden ist und hochqualitative Aufnahmen von Dialogen zu fingierten Terminabsprachen enthält: Mit vorgegebenen Kalendern wurden hypothetische Geschäftstreffen vereinbart. Durch die Laborsituation und die Vorgaben ist die Entstehungssituation der Gespräche nicht gänzlich natürlich oder spontan. Allerdings diente das entstandene Material einem klaren kommunikativen Ziel und ist damit authentisch. Die Aufnahmen wirken für die Situation geschäftlicher Telefondialoge sehr natürlich, auch wenn ihr Wortschatz speziell ist. Besonders in Abgrenzung zu gelesener Sprache zeigen sie die bekannten Merkmale wie Häsitationen und unvollständige Sätze, die spontan produzierte Sprache charakterisieren.

Material von 32 Sprechern (14 Frauen, 18 Männer) des Standarddeutschen mit meist norddeutscher Ausprägung wird zur Untersuchung herangezogen. Die Aufnahmen dauern etwa 3 3/4 Stunden (16 kHz, 16 bit). Dazu enthält das Korpus auch orthografische Umschriften, daraus generierte kanonische Aussprachen und auch phonetische Transkriptionen mit Segmentierung. Bei letzterer handelt es sich um eine allophonische Transkription (vgl. Vieregge, 1989) in *Sampa* (Gibbon et al., 1997)[1] mit einigen Zusätzen. Es wurden auf Basis phonemischer Symbole also auch weitere phonetische Informationen vermerkt: So die stellungsbedingten Unterschiede [ç] und [x]; Verschlussphase wird gegenüber Aspiration bei Plosiven unterschieden; Vokalsegmenten wird gegebenenfalls Nasalität und Laryngalisierung zugeordnet. /r/-Realisierungen sind mit Ausnahme von Vokalisierungen nicht unterschieden. Da sich bei der Annotation der Funktionswörter so weit wie möglich an die kanonische Aussprache gehalten wurde[2], ist der Status der lautlichen Segmente gegenüber Inhaltswörtern verschieden. Im Weiteren wird mit den Begriffen *Phon* und *Silbe* auf die transkribierten Symbole verwiesen, die mit ihrem Bezug zum Sprachsignal eindeutige zeitliche Grenzen und akustische Merkmale aufweisen. Es werden in der vorliegenden Arbeit sowohl die symbolisch umschriebenen Phone als auch die daraus ermittelten Silben als

[1] `http://www.phon.ucl.ac.uk/home/sampa/index.html`
[2] Ausnahmen bilden z. B. Elisionen.

Segmente bezeichnet. Der Bezug auf die Transkription stellt eine sinnvolle Operationalisierung für tatsächlich geäußerte Phone dar, während bei Gruppen mit gleichem Symbol doch besser von Phonen als Phonemen gesprochen wird, weshalb sie eckig geklammert dargestellt werden. Zum einen handelt es sich bei den Fragestellungen um die Phonemrealisierungen, zum anderen werden solche Realisierungen u. a. für betonte und unbetonte Silben unterschieden.

Für die Analysen wurde das Korpus in das Format der *Emu* Datenbank (Harrington und Cassidy, 2001) konvertiert, um dessen Eigenschaften für Datenabfragen nutzen zu können. Trotz des hilfreichen Tools *labconvert* von *Emu* war erhebliche händische Nacharbeit bei der Konvertierung notwendig, die knapp die Hälfte der Annotationsdateien betraf. Des Weiteren wurde eine Silbensegmentation vorgenommen, da diese noch nicht in der Annotation enthalten war und eine weitere Segmentebene eingeführt, in der Verschlussphasen und Aspirationen zu je einem Plosiv zusammengefasst wurden, um nicht in der Berechnung der PLSR von Pfitzinger (1999) abzuweichen. Auf dieser Basis wurde die PLSR berechnet und jeweils als eigenes Signal für jede Sprachdatei gespeichert.

8 Variation der Sprechgeschwindigkeit

In diesem und den folgenden Kapiteln werden Analyse und Interpretation der erhobenen Daten aus dem Kielkorpus dargestellt. Die Untersuchung der lokalen Sprechgeschwindigkeit im vorliegenden Kapitel dient dazu, einen Überblick über die Variationsbreite für das Korpus zu erhalten. Die Häufigkeitsverteilung von PLSR-Raten in der Mitte aller Silben,[1] ausgenommen Häsitationen, wird in Abbildung 8.1 dargestellt. Die wenigen deutlichen „Ausreißer" wurden überprüft und sind entweder auf extreme Segmentlängungen mit dem Charakter von Häsitationen oder kurze Silben meist aufgrund von nicht gekennzeichneten Aufnahmeschnitten und Signalfehlern zurückzuführen. Da die PLSR weitgehend normalverteilt ist, wird keine weitere Bereinigung der Daten vorgenommen, außer Daten im Abstand von 3 Standardabweichungen auszuschließen (148 Fälle: 0,27%), sodass fast 55000 Silben ausgewertet werden. Auf eine Transformation der Werte mit dem Logarithmus, wie bei Segmentdauern üblich, kann demnach bei der PLSR verzichtet werden.

Ein Vergleich zwischen PLSR und Silbendauern (Abb. 8.2) zeigt, dass eine Verkürzung der Silbendauern nicht gleichmäßig zu einem höheren wahrgenommenen Tempo führt: Gerade bei sehr niedrigem Tempo erhöht eine Verkürzung der Silbendauer die PLSR viel stärker als bei mittlerem oder hohem Tempo. Die Werte der PLSR mit einem Mittelwert um 100 entsprechen der Skala von Perzeptionstests, die in Pfitzinger (2001) erläutert werden. Für eine intuitivere Beurteilung werden sie von nun an transformiert dargestellt, um mit der gefensterten Silbenrate vergleichbar zu sein, die etwas stärker als die Phonrate gewichtet ist ($PLSR_{transf} = PLSR \cdot 0,06587 - 0,8$). Das neue Maß wird als $^{\text{Sil}'}/_{\text{s}}$ bezeichnet, da Silbenrate und PLSR nicht identisch sind (Abb. 8.3).

Um tempobedingte Aussprachevariationen analysieren zu können, muss zuvor die Verteilung der PLSR auf Systematiken untersucht werden. So können wichtige Faktoren identifiziert werden, die später zu berücksichtigen sind. Außerdem wird ein Überblick über Variation und Streuung des lokalen Tempos gewonnen. Eine bedeutende Frage ist die nach dem Tempoverhalten der Sprecher. Es

[1] Durch die Fensterung entspricht dies de facto dem mittleren Tempo in der Silbe ($r > .98$).

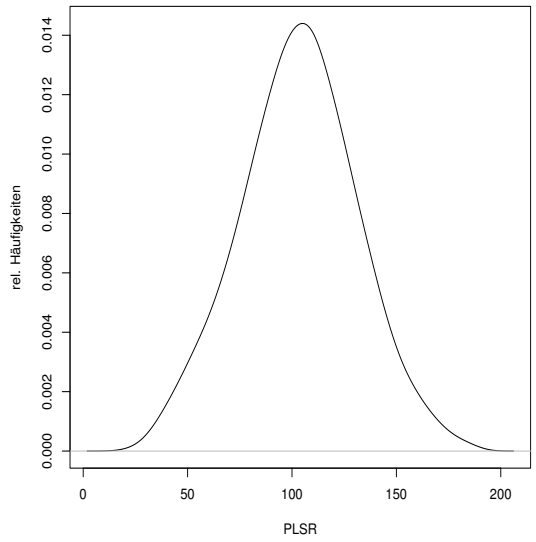

Abbildung 8.1: PLSR: Häufigkeitsverteilung, alle Sprecher

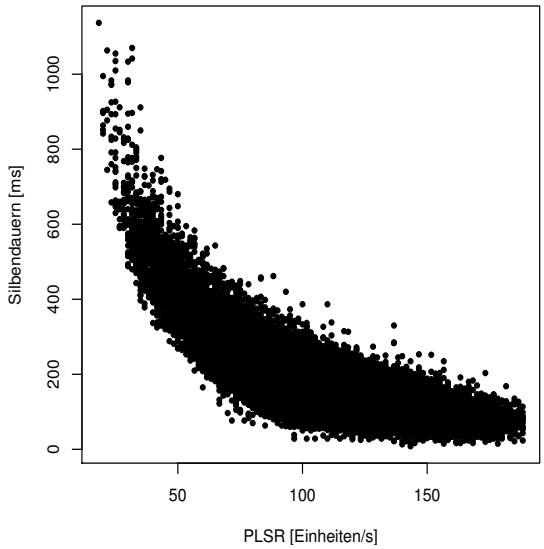

Abbildung 8.2: PLSR im Vergl. zu reziproken Silbendauern, alle Sprecher

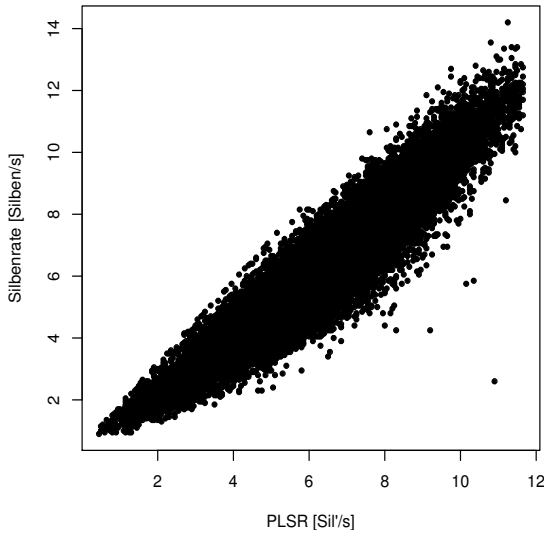

Abbildung 8.3: PLSR transformiert (in $^{\text{Sil}'}/_{\text{s}}$) gegenüber Silbenrate, alle Sprecher

gilt zu klären, ob die Sprecher sich in ihren Tempi stark unterscheiden und ggf. Gruppen zugeordnet werden können. Wie verhalten sich die Variationen innerhalb einzelner Sprecher gegenüber den Unterschieden zwischen ihnen? Ist das Geschlecht von Bedeutung?

8.1 Unterschiede zwischen den Sprechern

Weibliche Sprecher sind mit 5,8 $^{\text{Sil}'}/_{\text{s}}$ durchschnittlich etwas langsamer als männliche (6,2 $^{\text{Sil}'}/_{\text{s}}$). Wird nur das Geschlecht betrachtet, ist dieser Unterschied bei einem einfachen t-Test aufgrund der vielen Werte hoch-signifikant. In Anbetracht der großen Abweichungen zwischen Sprechern eines Geschlechts kann aber nicht mehr von einem signifikanten Effekt gesprochen werden ($F(1, 30) = 3.52$; $p = .07$).[2] Insofern bekräftigt dieses Ergebnis die Re-Analyse von Interviews im Niederländischen (Verhoeven et al., 2004) durch Quené (2005). Quené weist nach, dass signifikante Geschlechterunterschiede im Tempo auf eine unzureichende Methodik beruhen. Weder Geschlecht, noch Alter oder Dialektregion der Sprecher beeinflusst ihr mittleres Sprechtempo. Maßgeblich ist vor allem die Länge der Intonationsphrase, wobei ältere Sprecher kürzere Phrasen und damit ein langsameres Tempo produzieren.

Das mittlere Tempo der einzelnen Sprecher reicht von 5,0 $^{\text{Sil}'}/_{\text{s}}$ bis 7,2 $^{\text{Sil}'}/_{\text{s}}$. Die Standardabweichung ist im Vergleich dazu mit 1,4 $^{\text{Sil}'}/_{\text{s}}$ bis 2,1 $^{\text{Sil}'}/_{\text{s}}$ recht ge-

[2] **PLSR** ist abhängige, von **Geschlecht** unabhängige Variable, und **Sprecher** ist als Zufallsfaktor modelliert.

ring, sodass die signifikanten Unterschiede in den Mittelwerten zwischen den Sprechern motivieren, Tempounterschiede von langsamen gegenüber schnellen Sprechern mit einer Gruppenbildung unterscheiden zu können ($F(31, 54906) = 176.70$; $p < .001$). Die Varianz zwischen einzelnen Sprechern beträgt $0,28$, die innerhalb eines Sprechers ganze $3,12$, was später durch Einbeziehung weiterer erklärender Faktoren wie etwa Betonung reduziert werden wird. Dennoch bleibt festzustellen, dass vornehmlich Unterschiede im Tempo innerhalb der Sprecher auftreten, damit also lokale Variation deutlich stärker ist als die sprecherspezifische globale Sprechgeschwindigkeit.

Die jeweils 10 langsamsten und schnellsten Sprecher werden eigenen Gruppen zugewiesen. Die Mittelwerte der beiden Gruppen betragen $5{,}4\,\mathrm{Sil'/s}$ und $6{,}7\,\mathrm{Sil'/s}$, was einen signifikanten Unterschied darstellt ($F(1, 18) = 126.87$; $p < .001$, **Sprecher** ist Zufallsfaktor). Offenbar ist für Sprecher mit höherem Tempo auch die Standardabweichung (SD) etwas höher. Die Korrelation ist aber nur mittelstark ($r = .53$).

Bezieht man die Dialog-Situation mit ein, ergibt sich ein überraschendes Bild: Die Varianz zwischen zwei Sprechern eines Dialoges beträgt nur noch $0,05$, die zwischen den verschiedenen Gesprächen (also jeweils beide Gesprächspartner zusammen) $0,30$. Dies ist ein deutlicher Hinweis auf Annäherung im Tempo von beiden Sprechern eines Dialoges. Ob die Ursache tatsächlich in einer Angleichung des Tempos liegt, oder etwa in ähnlich langen Phrasen oder vergleichbarer Wortwahl und Ausdrücken, ist hier ohne zusätzliche Analysen nicht zu klären. Genauso bleibt offen, ob es einen dominanten Gesprächsteilnehmer gibt, oder beide Sprecher einander angleichen. Die mittleren Tempi der Gesprächspartner korrelieren miteinander jedenfalls mit $r = .94$ sehr stark (vgl. Abb. 8.4).

Da das Alter der Gesprächspartner eines Dialoges wenig voneinander abweicht ($SD = 1,9$ Jahre im Dialog gegenüber $SD = 5,0$ zwischen den Dialogen, bei maximaler Differenz von 40 Jahren zwischen dem Jüngsten und Ältesten), könnte auch ein Einfluss des Alters die Unterschiede der Mittelwerte zwischen den Dialogen erklären. Alte Sprecher sind durchschnittlich langsamer als jüngere (45–60 Jahre: $5,5\,\mathrm{Sil'/s}$, 20–35 Jahre: $6,2\,\mathrm{Sil'/s}$). Während die älteren Sprecher einen Median von $5,0$–$5,9\,\mathrm{Sil'/s}$ aufweisen, zeigen die jüngeren stärke Variabilität in den mittleren Tempi ($5,3$–$7,2$) (vgl. Abb. 8.5). Dieser Unterschied ist bei einem Mann-Whitney Test signifikant ($U(10, 22) = 16$; $p < .001$), zeigt aber nur eine mittlere Rangkorrelation nach Spearman ($\rho = -.58$).

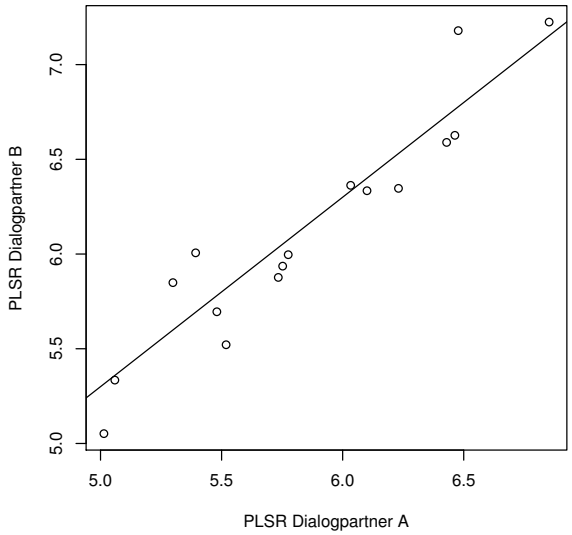

Abbildung 8.4: mittlere PLSR (in Sil′/s) von je beiden Gesprächspartnern „A“ und „B“, sowie Korrelationsgerade

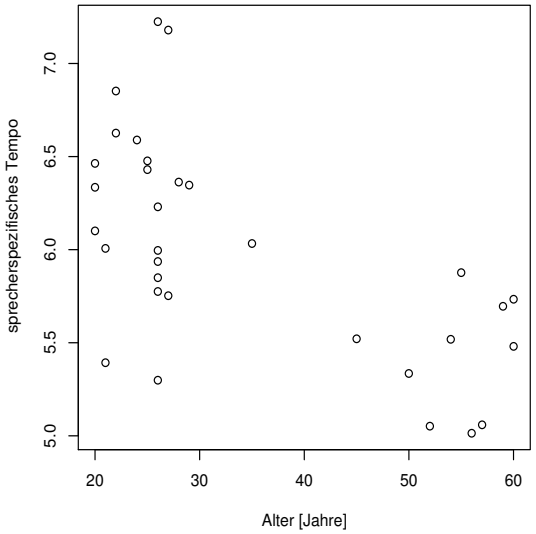

Abbildung 8.5: mittlere PLSR (in Sil′/s) der Sprecher gegenüber ihrem Alter

8.2 Systematische Variation in verschiedenen linguistischen Bedingungen

In diesem Kapitel wird zum einen beschrieben, wie lokales Tempo von Silben mit linguistischen Faktoren zusammenfällt. Zum anderen werden diese systematischen Unterschiede auf ihre Signifikanz überprüft. Berücksichtigt werden dabei **Wortbetonung** (hauptbetont, nebenbetont, unbetont; nur bei Inhaltswörtern), **Wortart** (Inhaltswort, Funktionswort), **Pausenumgebung** (Silbe grenzt direkt an Pause oder Häsitation), **Segmentanzahl** (variabel) und **Worthäufigkeit** (Anzahl von Wortformen sowie binäre Kategorie häufig/selten).

Informationen über die Phrase, wie etwa Phrasenlänge, werden nicht mit einbezogen, da sie nicht zugänglich sind und für die weitere Analyse von Aussprachevariationen von Wörtern nicht relevant genug erscheinen, um eine steigende Variablenanzahl zu rechtfertigen.

Da umfangreiche Wortfrequenzlisten des Deutschen – soweit dem Autor bekannt – ausschließlich auf Schriftsprache (Zeitungskorpora) basieren, die kommunikative Funktion der Terminplanung dagegen fachspezifisch ist, werden die Worthäufigkeiten aus dem Auftreten im Korpus ermittelt. Sie bilden damit eine mögliche Erwartungshaltung von Gesprächspartnern besser ab. Die Häufigkeit bezieht sich nicht auf das Lemma, sondern auf die Wortform. Dieses Maß ist sehr einfach zu extrahieren und wurde bereits erfolgreich für eine Analyse von Dauermaßen im Deutschen verwendet (Werner et al., 2003). Die Bedeutung dieses Faktors ist mit dem kontextsensitiver probalistischer Maße wie dem *Trigramm* zu vergleichen. Nur bei hochfrequenten Funktionswörtern korreliert das *Trigramm* stärker mit der Wortlänge als die Wortfrequenz (*Unigramm*) (Fosler-Lussier und Morgan, 1999). Dieser Unterschied ist jedoch für die empirische Untersuchung nicht relevant. Zusätzlich werden verschiedene Maße für die Segmentanzahl miteinander verglichen.

8.2.1 Linguistische Domäne lokalen Tempos

Wie segmentaler Kontext und Strukturen höherer linguistischer Ebenen auf Segmentdauern und damit auch auf lokales Tempo einwirken, wurde in Kapitel 2 ausführlich besprochen. In diesem Kapitel wird dazu die Frage bearbeitet, innerhalb welcher prosodischen Einheit *Timing* anscheinend für das vorliegende Kor-

pus organisiert ist. Innerhalb welcher Domäne sind also Segmentdauern weitgehend von ihrer Anzahl und kaum von anderen Faktoren (zum Beispiel von Betonung) abhängig: Mögliche Domänen umfassen unter anderem Silbe, Fuß oder Wort. Ein einfacher Ansatz zur Lösung dieser Frage ist, das Tempo über die Anzahl der Segmente innerhalb dieser Domänen zu erklären. Die prosodische Einheit mit der geringsten Varianz ist demnach die sinnvollste Wahl, da in ihren Grenzen das *Timing* gleichmäßiger bleibt als in anderen. Trotz fehlender artikulatorischer Evidenz ist dies ein aussagekräftiger Ansatz. Die Ergebnisse sind dabei vermutlich sprachspezifisch. Für das Britische erscheint die *narrow rythm unit* (NRU) – der Fuß bis zum Wortende – als beste Wahl, um Phondauern über ihre Anzahl zu erklären. Die unbetonten Silben an Wortanfängen (Anacrusis) zeigen starke Variabilität, während die Varianz in den Dauern innerhalb einer NRU deutlich geringer ist (Hirst und Bouzon, 2005). Für die Ebene des Wortes ergaben sich geringere Korrelationen zwischen Phonanzahl und Phondauer.

Dieses Ergebnis soll für das Deutsche überprüft werden: Dabei geht es nicht um Phondauern, sondern die PLSR, die als neues Maß für lokales Sprechtempo noch keiner solchen Analyse unterzogen wurde: Dazu wird die Abhängigkeit der PLSR von der Anzahl von Phonen und auch Silben in prosodischen Einheiten überprüft. Plosivaspiration, die im Korpus einem eigenen Segment zugeordnet wurden, werden aufgrund der zeitlichen Ausdehnung als eigenständiges Segment bei der Phonanzahl gewertet. Die Datenbasis für diese Auswertung basiert auf den PLSR-Werten für einzelne Silben, die auch in den bisherigen Analysen in diesem Kapitel verwendet wurden (vgl. S. 75). Eine Erhebung von PLSR für alle Phone ergibt wegen der langsamen Veränderung der PLSR innerhalb einer Silbe keinen zusätzlichen Informationswert. Von diesem Datensatz werden solche Silben ausgeschlossen, die zu Pausen und Häsitationen benachbart sind, um Einflüsse größerer prosodischer Einheiten, wie die Intonationsphrase, und Disfluenzen zu minimieren. Über 43500 Fälle gehen in die Analyse mit ein. Einflüsse globalen Tempos durch Sprecher oder Dialog sowie von intrinsischen Dauern vom Silbenkern sollen mit einer z-Transformation der Werte über **Sprecher** und **Silbenkern** minimiert werden.

Alle Variablen mit Ausnahme von Sprecher und Dialog (aufgrund der z-Transformation) zeigen bei einer Varianzanalyse hoch-signifikanten Einfluss auf die transformierten Geschwindigkeiten ($p < .001$): Alle Variablen für Segmentanzahlen, **Wortart** ($F(1, 43441) = 515.80$), **Betonung** ($F(2, 43441) = 593.66$), **Silbenkernidentität** ($F(42, 43441) = 7.86$) und **Worthäufigkeit** ($F(1, 43441) = 144.26$).

Mit der Transformation reduzieren sich die Unterschiede durch den Silbenkern. Da diese verschieden distribuiert sind, lässt sich ein signifikanter Einfluss nicht vermeiden.

Für die Auswahl der geeignetsten Korrelate für Segmentanzahl werden verschiedene Kombinationen über klassische lineare Modelle berechnet und die recht niedrigen R^2-Werte miteinander verglichen: Wird die Aspiration als eigenes Segment gewertet, kann Segmentanzahl die Tempodaten besser erklären. Dies ist in sofern überraschend, da Aspiration bei der Erhebung der PLSR nicht als eigenes Segment gewertet wird. Hier übt die Aspiration aufgrund der eigenen Dauer einen Einfluss auf auf die Silbenrate aus und könnte auch Veränderungen in den benachbarten Phondauern nach sich ziehen. Die Verbesserung durch Aspiration am Anteil der erklärten Varianz beträgt etwa 5% (relativ) für die Silbe auch in verschiedenen Bedingungen für Betonung, die hier gesondert behandelt werden (siehe Kapitel 8.2.3). Insgesamt ist das lokale Tempo bei betonten Inhaltswörtern besser vorherzusagen als bei unbetonten oder Funktionswörtern. Die Variabilität steigt um etwa 30% (relativ) für beide unbetonte Bedingungen.

Bei der Überprüfung aller erhobenen Korrelate für Segmente- und Silbenanzahl innerhalb einer prosodischen Domäne zeigten Phone in der Silbe den stärksten erklärenden Einfluss auf die PLSR. Dieser Einfluss ist auch stärker als als der von Segmenten in der NRU, obwohl dort ja nicht immer alle Phone und Silben mit einbezogen werden.

Die Anzahl von Silben im Fuß erklärt etwa 3% mehr Varianz als im Wort (10% bei unbetonten Silben in Inhaltswörtern).[3] Beim Vergleich der Kombination zwischen Phonen in der Silbe mit entweder Phonen oder Silben im Fuß erhöhen Silben im Fuß den R^2-Wert bei Inhaltswörtern (3–8% relativer Zuwachs). Bei Funktionswörtern sind die Ergebnisse nicht verschieden. Insgesamt erklären die Phone in der Silbe das lokale Tempo etwas stärker als Silben im Fuß. Die Kombination beider Maße bringt eine weitere Steigerung auf R^2-Werte um etwa 30%, da sie recht unabhängig von einander sind. Das Wort als Domäne zeigte durchweg niedrigere R^2-Werte.

Da bei mehr Phonen in der Silbe das lokale Tempo sinkt, dagegen mehr Silben im Fuß die PLSR steigen lassen, ergänzen sich beide prosodische Domänen für

[3] Die R^2-Werte sind äußerst niedrig, da hier aufgrund der Fragestellung nicht intrinsische Tempi/Dauern von Segmenten miteinbezogen sind, mit denen dann mittlere Regressionen erreicht werden. Soweit sind die sind die hier erreichten Werte durchaus mit anderen Untersuchungen vergleichbar (etwa Tyson, 2005).

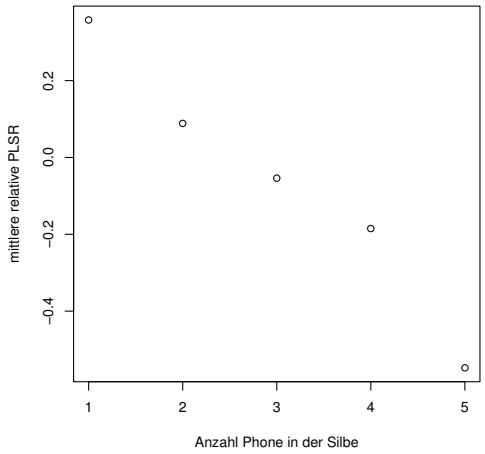

Abbildung 8.6: Phonanzahl und PLSR, z-transformiert und über die Bedingungen betonte Inhaltswörter, unbetonte Inhaltswörter und Funktionswörter gemittelt

Timing im Deutschen. Innerhalb eines Fußes wird mit steigender Silbenanzahl zumindest perzeptiv (geringere PLSR) das Tempo erhöht. Die tatsächlichen Dauern der Füße werden natürlich nur relativ zur Silbenanzahl verringert, während sie tatsächlich länger werden. Dagegen wirkt eine Silbe langsamer (PLSR), wenn sie mehr Phone enthält, was sich auch schon aus der etwas stärkeren Gewichtung in der Definition von PLSR ableiten lässt.

Wie weiter oben dargestellt, sind die Haupteffekte für Segmentanzahl und PLSR signifikant. Im Folgenden werden genauere Varianzanalysen anhand der z-transformierten Werte vorgenommen, und für signifikante Unterschiede Erwartungswerte mittels **nlme** aus den untransformierten Daten erhoben.

8.2.2 Segmentanzahl

Je mehr Phone in der Silbe sind, desto niedriger wird die PLSR. Obwohl die Werte für PLSR enorm streuen, werden sie mit einer Regressionsgraden modelliert, um beide Variablen nicht kategorisieren zu müssen. Die Nebeneffekte der Signifikanztests betreffen die Betonung. Innerhalb verschiedener Bedingungen (betont und unbetont in Inhaltswörtern, Silben in Funktionswörtern) sinken die Mittelwerte der PLSR nach grafischer Überprüfung stetig und weitgehend linear mit steigender Anzahl von Phonen. Einzige Ausnahme bilden hauptbetonte Silben mit 3 und 4 Phonen pro Silbe, deren Verteilung und Mittelwert nahezu identisch sind. In Abbildung 8.6 sind die gemittelten z-transformierten Werte der PLSR zur Anzahl von Phonen in der Silbe aufgetragen. Durchschnittlich

nimmt die PLSR pro Phon mehr in der Silbe um 0,41 $^{\text{Sil}'}/_{\text{s}}$ bei Funktionswörtern ab; 0,36 $^{\text{Sil}'}/_{\text{s}}$ bei unbetonten Inhaltswörtern und 0,33 $^{\text{Sil}'}/_{\text{s}}$ bei betonten.

Je mehr Silben im Fuß, desto höher wird das Tempo, pro Silbe etwa 0,22 $^{\text{Sil}'}/_{\text{s}}$ (0,18 für unbetonte Silben in Inhaltswörtern; 0,24 für betonte Silben).

8.2.3 Betonung

Unterschiede in der Betonung werden nur bei Inhaltswörtern untersucht, da bei Funktionswörtern nur in Ausnahmefällen Wortakzente realisiert werden. Betonte Silben in Inhaltswörter weichen signifikant von unbetonten ab ($F(2, 25661) = 24.72$; $p < .001$). Im Vergleich zu unbetonten mit etwa 7 $^{\text{Sil}'}/_{\text{s}}$ zeigen dabei nebenbetonte Silben mit etwa -0,42 $^{\text{Sil}'}/_{\text{s}}$ einen wesentlich größeren Unterschied als hauptbetonte gegenüber unbetonten mit -0,21 $^{\text{Sil}'}/_{\text{s}}$. Diese Differenz zwischen haupt- und nebenbetonten Silben ($F(1, 14351) = 44.13$; $p < .001$) liegt allerdings an der ungleichen Verteilung von Phonen pro Silbe, die sich durch einen Nebeneffekt auch zeigt ($F(1, 14249) = 45.11$; $p < .001$). Im direkten Vergleich von Silben in Wörtern mit Haupt- und Nebenbetonung (2306 Wortrealisierungen) bestehen hauptbetonte Silben vor allem aus 2 und 3 Segmenten, dagegen nebenbetonte aus 2–5.[4] So sind auch die Vokaldauern in diesen Silben kürzer, ihre Silbendauern aber länger.

Um zu überprüfen, ob haupt- und nebenbetonte Silben sich auch bei gleicher Segmentanzahl in ihrem lokalen Tempo unterscheiden, werden Mann-Whitney Tests durchgeführt. Für die Population von Wörtern mit Nebenbetonungen ergeben sich keine signifikanten Unterschiede in PLSR zwischen Haupt- und Nebenbetonung, wenn jeweils Silben mit gleicher Phonanzahl (zwei[5] oder vier[6]) miteinander verglichen werden. Die Ausnahme bilden Silben mit drei Segmenten, wo nebenbetonte eine geringere PLSR aufweisen als hauptbetonte ($U(676, 481) = 177956$; $p < .01$). Bei diesen Statistiken wäre eine weitere Unterteilung nach der Anzahl von Silben im Wort sowie Identität des Silbenkerns zweckmäßig. Dies würde allerdings zu nicht vergleichbaren Populationen führen. Insofern lässt das signifikante Ergebnis keine Interpretation dahingehend zu, dass nebenbetonte Silben tatsächlich langsamer seien als hauptbetonte.

Als Folge werden alle folgenden Hypothesen jeweils mit getrennten Datensät-

[4] Beispiele sind Zahlwörter wie „dreiundzwanzigster“ oder „einverstanden“.
[5] $U(509, 1127) = 284327$; $p = .61$
[6] $U(515, 214) = 56010$; $p = .36$

zen für betonte und unbetonte Inhaltswörter überprüft. Die nebenbetonten Fälle werden aufgrund ihrer geringen Anzahl und damit auffälligen Verteilung ausgeschlossen. Ebenso werden auch unbetonte Silben in Inhalts- gegenüber Funktionswörtern unterschieden.

8.2.4 Wortart

Da bereits der große Einfluss der Silbenanzahl dargestellt wurde, macht eine über eine Beschreibung hinausgehende Analyse für verschiedene Wortarten nur Sinn bei gleicher Silbenanzahl. Funktionswörter sind vor allem Ein- und Zweisilber, unbetonte Silben in Inhaltswörtern betreffen mindestens Zweisilber. Bei Wörtern mit zwei Silben sind unbetonte Silben in Funktionswörtern 0,69 $^{\text{Sil}'}/_{\text{s}}$ schneller als solche in Inhaltswörtern ($F(1, 8957) = 605.30$; $p < .001$, Kontrollvariable: **Phonanzahl in Silbe** (1–4) als Faktor). Dieser Unterschied ist abhängig von der Anzahl der Phone in der Silbe ($F(3, 8957) = 9.36$; $p < .001$, für die Interaktion beider Faktoren): 0,88; 0,86 und 0,46 $^{\text{Sil}'}/_{\text{s}}$ für 1–3 Segmente (in Post-Hoc Vergleiche jeweils $p < .001$), während es keinen Unterschied bei vieren gibt ($p = .42$).

Dass selbst Silben in Funktionswörtern im Vergleich zu unbetonten Silben noch signifikant schneller sind, belegt, wie wichtig eine Trennung nach Wortart sogar für nur zwei Gruppen ist. Diese Trennung auch für nicht-betonte Silben zwischen Inhalts- und Funktionswörtern wird in den weiteren Analysen beibehalten.

8.2.5 Wortfrequenz

Da im vorliegenden Datensatz kaum ein Inhaltswort besonders häufig vorkommt, wird von einer Analyse von Inhaltswörtern nach Zusammenhängen zwischen lokalem Tempo und Wortfrequenz abgesehen. Deshalb werden hier nur Funktionswörter verglichen, und aufgrund der Verteilung nur Einsilber, da häufige Wörter in Zweisilbern nicht ausreichend oft auftreten: Allgemein sind Silben in häufigen Funktionswörtern schneller ($F(1, 13361) = 51.21$; $p < .001$). Der Effekt von Wortfrequenz interagiert mit der Anzahl Phone in der Silbe ($F(3, 13361) = 10.66$; $p < .001$). Post-Hoc Vergleiche zeigen, dass die Wortfrequenz nur bei zwei Phonen ($p < .001$) in der Silbe ($0,50\,^{\text{Sil}'}/_{\text{s}}$) signifikant ist.

Selbst bei den seltenen Funktionswörtern treten keine außergewöhnlichen Wör-

ter auf. Da die Wortfrequenz einzig in einer Position signifikant auf das lokale Tempo einwirkt, wird sie nicht als Variable bei den Ausspracheuntersuchungen verwendet.

8.2.6 Silbenkern

Um den Einfluss der Identität des Silbenkerns auf das Tempo der Silbe zu untersuchen, werden die Daten nur für die Sprecher z-transformiert. Für diese variiert PLSR signifikant für verschiedene Phon-Klassen in Funktionswörtern ($F(36, 17588) = 24.82$; $p < .001$), unbetonten Inhaltswörtern ($F(30, 8229) = 12.54$; $p < .001$) und betonten ($F(37, 9615) = 21.32$; $p < .001$). Eine Aufstellung einzelner Phonklassen wäre nicht aussagekräftig, da aufgrund des Charakters der PLSR dann auch die Konsonanten in der Silbe berücksichtigt werden müssten. Es wird aber eine Analyse nach einigen distinktiven Merkmalen vorgenommen:

Unbetonte Silben mit sogenannten Kurzvokalen (außer [ə][7] und [ɐ]) sind im Mittel 0,2 $^{\text{Sil}'}/_{\text{s}}$ schneller als unbetonte Silben mit Langvokalen in Inhaltswörtern ($F(1, 3133) = 68.60$; $p < .001$). In betonten Silben sind Kurzvokale 0,56 $^{\text{Sil}'}/_{\text{s}}$ schneller ($F(1, 5096) = 143.59$; $p < .001$). Bei Funktionswörtern ergibt sich ein Tempounterschied von 0,29 $^{\text{Sil}'}/_{\text{s}}$ ($F(1, 11856) = 37.97$; $p < .001$). Eine Überprüfung für Minimalpaare mit gleichen artikulatorischen Merkmalen außer der Länge (u. a. [a], [aː] / [ɛ], [eː]) bestätigt dieses Verhalten.

Bei der Zungenlage wird nur **vorn** gegenüber **hinten** überprüft, da Phon-Klassen mit mittlerer Lage entweder von ihrem Auftreten im Wort beschränkt sind ([ə], [ɐ]) oder als [a], [aː] im Deutschen keine Vergleichsklasse mit tiefer Zungenlage aufweisen. Vordere Vokale sind in Funktionswörtern 0,29 $^{\text{Sil}'}/_{\text{s}}$ schneller als hintere ($F(1, 6590) = 30.03$; $p < .001$). In Inhaltswörtern sind vordere Kurzvokale nur in unbetonten Silben mit 0,47 $^{\text{Sil}'}/_{\text{s}}$ signifikant schneller ($F(1, 964) = 27.00$; $p < .001$).

Die Zungenhöhe eines Vokals ist auch abhängig von der Bewegung des Kiefers und damit potentiell mit Tempo korreliert. Auch wenn zentrale Zungenhöhen deutlich schneller sind als Vokale mit tiefen oder auch hohen Zungenhöhen, zei-

[7] Die segmentierten und annotierten Phone im Korpus werden hier durchweg als Gruppen von Phonen angesehen und mit eckigen Klammern dargestellt. Zumindest Phone in Funktionswörtern, die ja über ein Aussprachewörterbuch annotiert wurden, könnten jedoch als Gruppen in den Analysen auch als Phoneme betrachtet werden. Hier stehen allerdings ihre Ausprägungen und damit ihre Realisierungen im Vordergrund und die Ergebnisse lassen sich nicht auf andere als den hier untersuchten Sprechstil übertragen.

gen Einzelanalysen kaum signifikante Ergebnisse. Die Tempounterschiede liegen demnach an der Verteilung der Phonklassen in den berücksichtigten Bedingungen. So sind vor allem ([ə], [ɐ]) für das hohe Tempo verantwortlich. Diese lassen sich aber nicht mit anderen Vokalen vergleichen. In Funktionswörtern ($F(3, 13379) = 27,32$; $p < .001$) und bei betonten Silben in Inhaltswörtern ergeben sich bei Ausschluss von [aː], [a], die langsamer sind als andere Zungenhöhen, keine signifikanten Unterschiede. In unbetonten Silben sind Hochzungenvokale im Mittel 0,25 $^{\text{Sil}'}/_{\text{s}}$ langsamer als andere ($F(3, 5952) = 6.89$; $p < .001$).[8] Dieses Ergebnis entspricht nicht anderen Studien (vgl. Kapitel 2), lässt sich aber durchaus dadurch erklären, dass hohe Zungenhöhen durchschnittlich eine längere Artikulationsbewegung mit sich bringen, da sie eine Extremstellung beim Sprechen darstellen.

Eine Überprüfung des Merkmals Rundung wird nicht vorgenommen. [øː], [yː] und [œ] treten nicht oft genug in vergleichbaren Bedingungen mit ungerundeten Partnern auf, sodass sich ein möglicher [ʏ],[ɪ] Unterschied nicht generalisieren ließe.

Die silbischen Konsonanten, die fast ausschließlich in unbetonten Silben auftreten, unterscheiden sich in ihrem Tempo nicht signifikant von Kurzvokalen.

8.2.7 Pausenumgebung

Z-transformierte Werte für Sprecher und Silbenkern werden auf ihr Tempo in der Umgebung von Pausen untersucht. Dabei richtet sich die Unterscheidung nach der Transliteration. Der Umschrift ist nicht zu entnehmen, ob eine stille oder gefüllte Pause geplant war oder eine solche Pause über häsititionalen Charakter verfügt. Es wird allerdings angenommen, dass stille Pausen in der Regel vorbereitet sind und der Strukturierung und dem Atmen dienen, während gefüllte Pausen öfter ungeplant sind (Butterworth, 1980), weswegen sie hier als Häsitationen bezeichnet werden. Schließlich kann ein Schlucken oder ungeplantes Atmen die Ursache für die gefüllte Pause sein, und Häsitationen signalisieren häufig während eines kognitiven Prozesses, dass gleich weitergesprochen wird. Dies ist ansonsten nicht notwendig, da ja Pausen zur Strukturierung erwartet werden.

In einer Studie mit vergleichbaren Daten aus dem Verbmobil-Projekt treten al-

[8] Diese wurden über Post-Hoc Vergleiche ermittelt.

lerdings 9/10 der gefüllten Pausen an syntaktischen Grenzen auf (Batliner et al., 1995). Dies ist Hinweis darauf, dass syntaktische Phrasen Planungseinheiten darstellen, an deren Grenzen durch kognitive Arbeit Pausen auftreten, die keine kommunikative Funktion haben, und daher mit „Häsitationen“ vom Sprecher entschuldigt werden. In Lesesprache treten dagegen Pausen vor allem bei Satz- und Teilsatzgrenzen auf (Deese, 1980).

Es werden nur Silben mit häufigen Vokalklassen verwendet, da so der Einfluss verschiedener Bedingungen auf die z-Transformation minimiert wird.

Silben vor einer stillen Pause sind deutlich langsamer als solche ohne unmittelbar benachbarte Pausen. Der Unterschied von ganzen $3,2\,\mathrm{Sil'/s}$ vor Pausen ist für Funktionswörter signifikant ($F(1,15350) = 2983.40$; $p < .001$), bei unbetonten Silben in Inhaltswörtern beträgt die Verlangsamung $2{,}4\,\mathrm{Sil'/s}$ ($F(1,7982) = 2284.36$; $p < .001$), bei betonten $2{,}0\,\mathrm{Sil'/s}$ ($F(1,3609) = 703.21$; $p < .001$). Vor Häsitationen sind Tempounterschiede nicht signifikant. Dies ist insofern bedeutsam, als dass der ungeplante Charakter von Häsitationen eine Erklärung für die fehlende Verlangsamung darstellen kann.

Silben nach Pausen oder Häsitationen verhalten anders als solche davor: Für Funktionswörter sind Silben nach Häsitationen $2{,}4\,\mathrm{Sil'/s}$ langsamer als solche ohne direkte Nachbarschaft ($F(1,10230) = 151.84$; $p < .001$). Betonte Silben in Inhaltswörtern sind $1{,}2\,\mathrm{Sil'/s}$ schneller nach Häsitationen als andere ($F(1,6383) = 40.82$; $p < .001$). Unbetonte Silben zeigen weder nach Pausen noch nach Häsitationen einen signifikanten Effekt.

Insgesamt ist die Variabilität von Tempo in Silben in direkter Umgebung von Pausen und Häsitationen mit über dem 3-fachen weit größer als anderweitig.

8.2.8 Zusammenfassung

Im vorliegenden Kapitel wurde das Auftreten und die Verteilung von lokaler Sprechgeschwindigkeit auf Silbenebene beschrieben und analysiert. Die Darstellung von Temposchwankungen und -verteilung im untersuchten Korpus ist notwendig für die weitere Analyse auf Aussprachevariationen, da nun sichergestellt ist, dass die Sprecher sich deutlich in ihrem globalen Tempo (mit bis zu $2\,\mathrm{Sil'/s}$) unterscheiden, und auch innerhalb der Sprecher und zwischen verschiedenen linguistischen Bedingungen das lokale Tempo signifikant variiert. Somit

sind trotz eng begrenzter Kommunikationssituation und Register so deutliche Schwankungen in der Sprechgeschwindigkeit vorhanden, dass diese auf ihren Zusammenhang mit Aussprachevariationen untersucht werden können.

Dabei zeigt sich, dass die berücksichtigten Bedingungen wie Betonung und Wortart mit in die weitere Analyse einfließen müssen. Dialogpartner weisen auffällig ähnliche mittlere Tempi auf: Ein starkes Indiz für eine Annäherung im Gespräch. Ältere Sprecher (die Grenze verläuft hier bei etwa 40 Jahren) zeigen dabei ein niedriges mittleres Tempo, unabhängig vom Geschlecht, während die jüngeren in diesem Wert stark schwanken, häufiger aber schneller sprechen.

Die Schwierigkeit, für die Hypothesentests vergleichbare Bedingungen zu finden, verdeutlicht das Zusammenwirken verschiedener linguistischer Ebenen bei der Verteilung von PLSR in den Silben: So weisen gerade die langsameren Silben Vokale aus Phon-Klassen auf, die in den langsamen betonten Silben in Inhaltswörtern auftreten, und genauso lassen sich Wortfrequenzunterschiede nicht von der Wortart trennen.

Dies bedeutet eine sehr robuste Verteilung linguistischer Informationen über verschiedene Ebenen hinweg, die sich im Tempoverhalten wiederfinden, und damit das Auffinden ursächlicher Faktoren bei diesen Daten unmöglich machen. Selbst das Einbinden größerer prosodischer Domänen wie Intonationsphrase oder Äußerung würde wieder mit dem Auftreten von Wortarten korrelieren. Lokales Tempo scheint hier sehr stark von phonetisch/phonologisch segmentalen Bedingungen bestimmt zu sein.

Im Korpus ist eine große Variabilität für das Tempo innerhalb der untersuchten Bedingungen wie Betonung und Wortart vorhanden, die weitgehend bei dem dreifachen der Unterschiede zwischen den Gruppen liegt. Hier sind weitere Faktoren anzunehmen, die lokales Tempo erklären können, beispielsweise Phrasenanfänge und -enden, aber auch genauere Aspekte von Diskontinuitäten und Silbenposition im Wort. Für solche Untersuchungen ist ein großes Korpus nur bedingt verwendbar. Zwar treten hier natürlich produzierte Tempovariationen und authentische Distributionen auf, und lassen so einheitliche Tendenzen zum Vorschein kommen. Aber feinere phonetische Details sind nur bei genauen Einzelanalysen überprüfbar, die streng kontrollierter Daten bedürfen. So zeigt der Faktor **Stimmhaftigkeit** beim vorliegenden Material weder vor dem Silbenkern, noch danach einen signifikanten Haupteffekt, weshalb diese Analyse auch nicht beschrieben wurde. Da die in diesem Kapitel vorgestellten Analysen nur

eine Grundlage für die jetzt folgenden bilden, können nicht alle möglichen Faktoren berücksichtigt werden. Vor allem ist der für Stimmhaftigkeit so wichtige phonetischen Kontext nicht ausgewogen und hätte genau kontrolliert werden müssen, sodass ein fehlender Effekt wenig aussagekräftig ist.

Es ist auch deutlich geworden, dass PLSR nicht vom phonetischen Material unabhängig ist, was zu besonderer Umsicht bei weiteren Untersuchungen führt, etwa die Kontrolle von Ergebnissen spektraler Reduktion bei Phon-Klassen durch zusätzliche Untersuchungen für einzelne Wörter, um so das phonetische Material zu kontrollieren.

9 Spektrale Analyse der Monophthonge

Im vorliegenden Kapitel werden spektrale Eigenschaften von Monophthongen in Bezug auf die Sprechgeschwindigkeit untersucht. Ausgeschlossen werden alle geäußerten Vokale mit direkter Pausenumgebung, um Artefakte durch diese besondere Stellung zu vermeiden, nasalierte und laryngalisierte Fälle wegen den problematischen akustischen Eigenschaften, sowie zu gering vertretene Vokal-Klassen. Die statistischen Tests werden jeweils getrennt nach betonten und unbetonten Inhaltswörtern, sowie Funktionswörtern durchgeführt. Genaue Ergebnisse der statistischen Auswertung befinden sich im Anhang A.1. Über 25000 Fälle gehen in die Analyse ein.

Die Werte für die beiden ersten Formanten werden mit LPC (Praat) ermittelt und über *Emu* (Harrington und Cassidy, 2001) – basierend auf einem einfachen Targetmodell – als Durchschnittswert bei 40%, 50% und 60% der Vokaldauer erhoben. Die Mittelung dient dem Ausgleich von Diskontinuitäten und der Minimierung des Einflusses konsonantischen Kontexts. Diese Wahl beruht nicht auf dem Vorzug einer Perzeptionstheorie, sondern ist sehr praktikabel. Die Analyseergebnisse können unabhängig von konkurrierenden Theorien verwendet werden, wie Pitermann (2000) zeigt. Obwohl phonetischer Kontext sicherlich noch einen Einfluss auf die mittleren Werte hat, werden Informationen über benachbarte Konsonanten nicht erhoben, da dies zu zu vielen Faktoren in den Statistiken führen würde. Die Formantwerte werden in der Einheit Bark (Traunmüller, 1989) untersucht, um der menschlichen Tonhöhenwahrnehmung gerecht zu werden.

Drei grundsätzliche Fragen sollen in diesem Kapitel beantwortet werden:

1. *Wenn ein Sprecher sein Tempo variiert, verändern sich dann auch die Formantfrequenzen seiner geäußerten Monophthonge?*

   Dazu werden sowohl die Formantwerte als auch das lokale Tempo z-transformiert, um Sprecherunterschiede (vor allem durch Geschlecht und intrinsisches Tempo) auszugleichen. Es wird also das relative Tempo mit relati-

ven Formantveränderungen verglichen (Kapitel 9.1).

2. *Sind schnelle gegenüber langsamen Sprechern in ihren Formantwerten verschieden?*

   Die Daten bei jeweils mittleren Tempi werden auf Unterschiede zwischen Sprechergruppen untersucht (Kapitel 9.2).

3. *Wenn sich Veränderungen ergeben, handelt es sich um Reduktionen oder verstärkte Koartikulation?*

   Diese Frage wird in Kapitel 9.3 behandelt.

Bei den vorliegenden Daten weist das Sprechtempo innerhalb der einzelnen Sprecher für jede Gruppe von Monophthongen eine Standardabweichung von 1,35 $^{\text{Sil}'}/_{\text{s}}$ auf. Dagegen beträgt die mittlere SD zwischen den Sprechern innerhalb dieser Gruppen nur 0,81 $^{\text{Sil}'}/_{\text{s}}$. Daher werden für die Analyse von Sprecherunterschieden nicht alle Personen berücksichtigt, sondern nur die langsamen und schnellen, wie in Kapitel 8.1 dargestellt. Die mittlere Differenz dieser beiden Gruppen (jeweils für verschiedene Bedingungen) ist 1,22 $^{\text{Sil}'}/_{\text{s}}$, bei einer mittleren SD innerhalb der Sprechergruppen von 0,64 $^{\text{Sil}'}/_{\text{s}}$.

Die Variabilität in PLSR hat sich durch Gruppierung nach Betonung, Wortart und Vokal innerhalb der Sprecher um über 40% verringert (vgl. Kapitel 8.1), sodass nicht nur grundsätzliche Überlegungen, sondern auch die Daten selbst zeigen, dass Formantanalysen für diese Bedingungen getrennt durchgeführt werden müssen.

9.1 Akustische Variation und relatives Tempo

Die Regressionsanalysen für normalisiertes Tempo zeigen hoch-signifikante Veränderungen in den relativen Formantfrequenzen bei relativen Temposchwankungen. Die signifikanten Resultate betreffen vornehmlich betonte Silben. Ausnahme hiervon sind u. a. für F_1 die Phonklassen [aː], [a], [ɔ] und für F_2 [eː], die auch bei unbetonten Silben und in Funktionswörtern signifikante Unterschiede aufweisen. Keine signifikanten Ergebnisse treten für [ɪ], [ʊ], [ʏ], [øː], [yː] auf, wobei für die letzten beiden Phonklassen kaum Fälle vorliegen.

Eine Illustration der Ergebnisse zeigen die Abbildungen 9.1 bis 9.6. Die Anno-

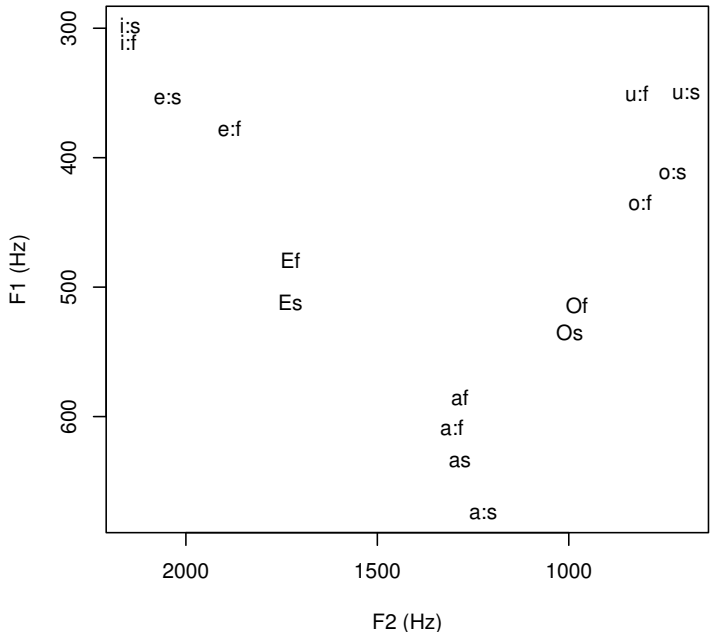

Abbildung 9.1: betonte Vokale in Inhaltswörtern (männliche Sprecher)

tation erfolgt aus technischen Gründen in SAMPA. Sichtbar sind jeweils nur signifikante Ergebnisse, mit den Mittelwerten der Formantfrequenzen (von Mittelwerten der Sprecher) für Männer und Frauen in Hz. Die Einteilung in schnelle (Vokal+f) und langsame Fälle (Vokal+s) erfolgt über die z-transformierten Werte für das Tempo einzelner Sprecher mit den Grenzen bei +-1 SD. Damit stellen die Grafiken keine Extremwerte dar, sondern verdeutlichen die Tendenzen der signifikanten statistischen Ergebnisse. Das [œ][1] ist mit seiner zentralen Stellung auffallend von theoretischen Werten entfernt. Dabei handelt es sich wohl um eine typische Reduktion im Deutschen für spontanen Sprechstil.

Die Veränderungen der Werte für höheres gegenüber niedrigerem Tempo lassen sich weitgehend als Zentralisierungen bezeichnen. Ausnahmen bilden die beiden zentralen unbetonte Vokale [ə][2] und [ɐ][3], die mit ansteigendem Tempo eine Verschiebung von F_1 zu niedrigeren Werten aufweisen (siehe Anhang A.1 für die genauen Ergebnisse). Des Weiteren erreicht beim [ɛ][4] F_1 für männliche Sprecher Werte unter 500 Hz. Dies ließe sich noch mit einer Tendenz zum Schwa anstatt zum hypothetischen Zentrum des Vokalraums erklären, da das Schwa im Deutschen in der Regel einen niedrigeren F1 aufweist (etwa 400 – 450 Hz). Dies würde dementsprechend auch für die Lage schneller [ə] und [ɐ] vom 1. For-

[1] In den Abbildungen als „9“ dargestellt (SAMPA).
[2] In den Abbildungen als „@“ dargestellt (SAMPA).
[3] In den Abbildungen als „6“ dargestellt (SAMPA).
[4] In den Abbildungen als „E“ dargestellt (SAMPA).

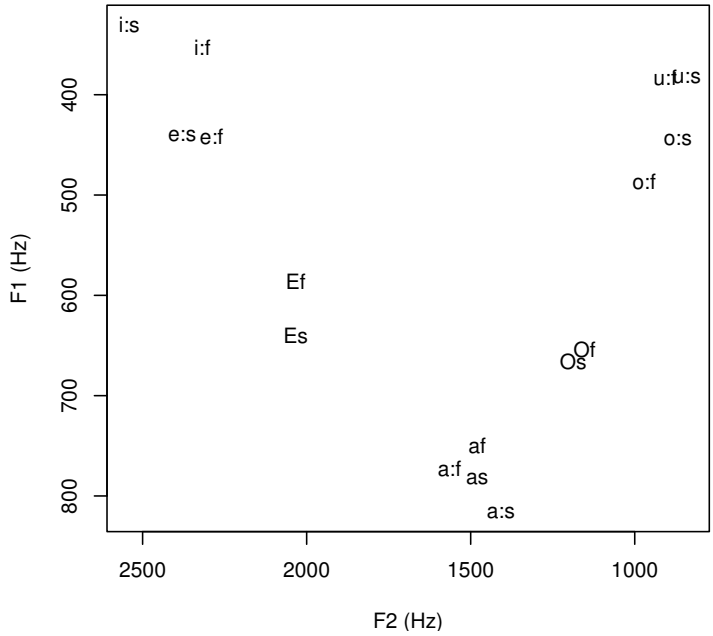

Abbildung 9.2: betonte Vokale in Inhaltswörtern (weibliche Sprecher)

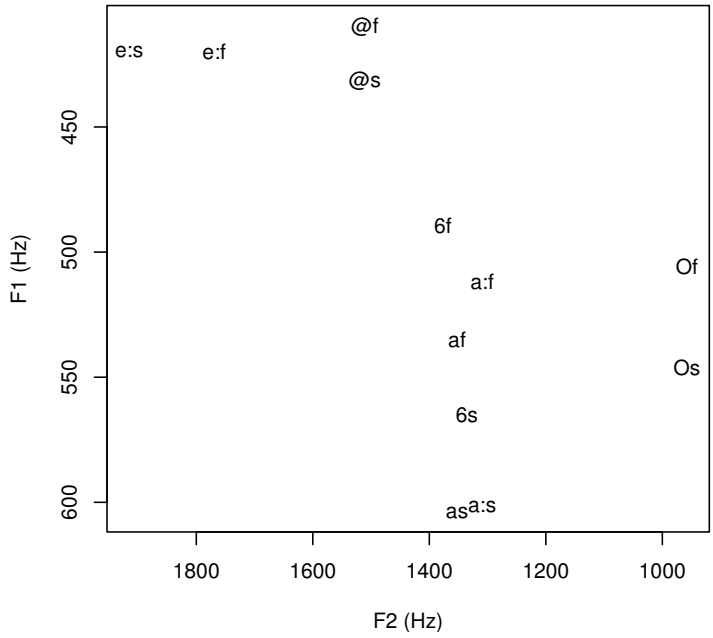

Abbildung 9.3: unbetonte Vokale in Inhaltswörtern (männliche Sprecher)

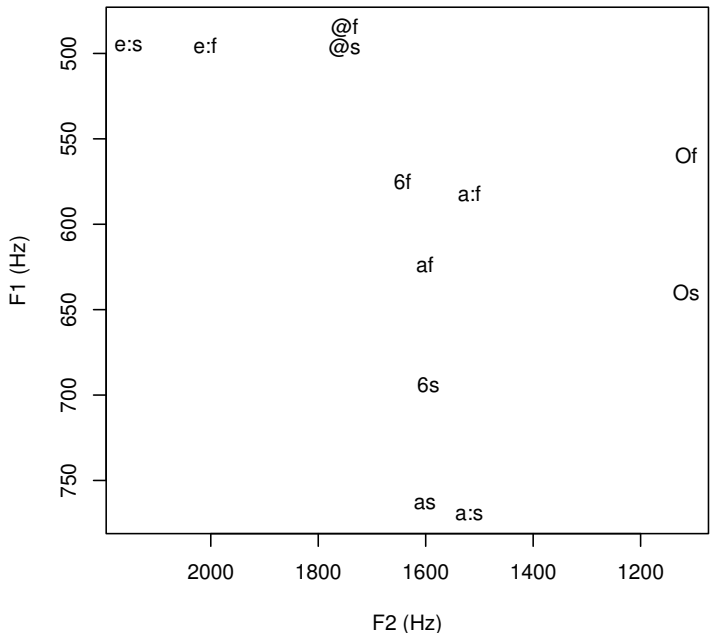

Abbildung 9.4: unbetonte Vokale in Inhaltswörtern (weibliche Sprecher)

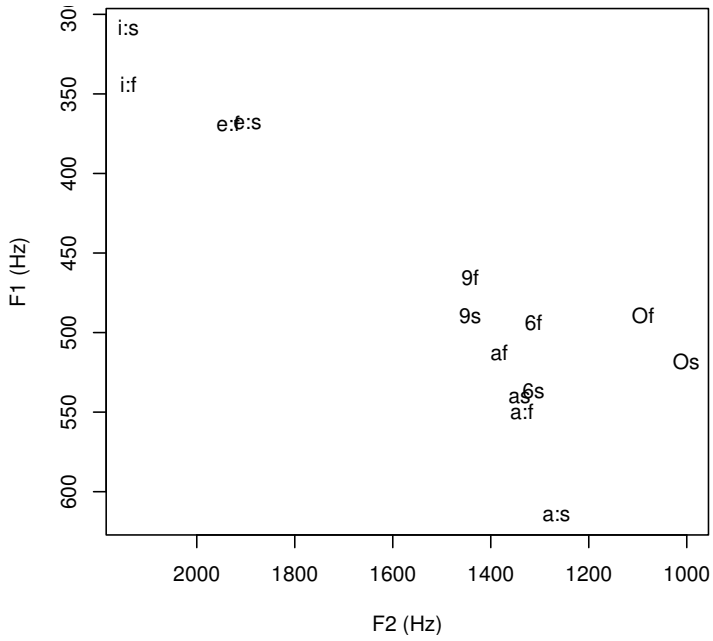

Abbildung 9.5: Vokale in Funktionswörtern (männliche Sprecher)

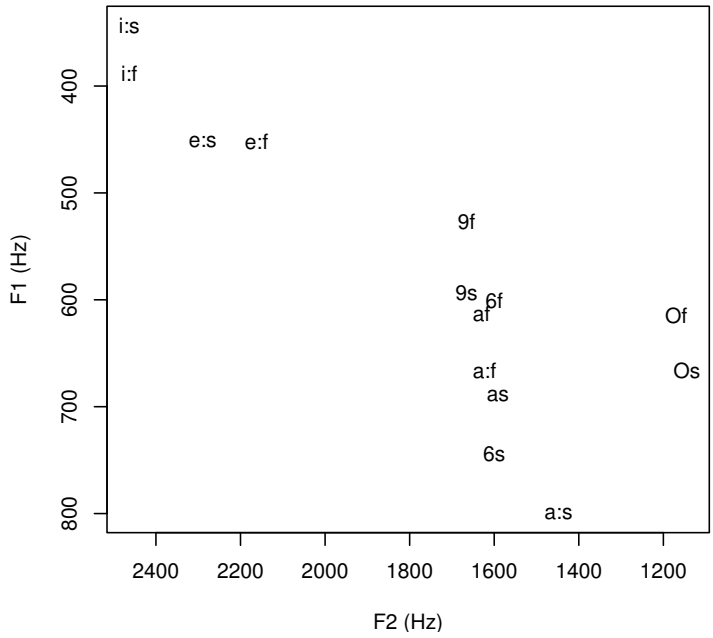

Abbildung 9.6: Vokale in Funktionswörtern (weibliche Sprecher)

manten gelten, die niedrigere Werte aufweisen als das hypothetische Zentrum des Vokalraumes von etwa 500 Hz für Männer und 550 Hz für Frauen. Beim [ɔ][5] verringert sich F_2 für beide Geschlechter bei betonten Silben. Für diese Dezentralisierung ergibt sich erst einmal keine Erklärung, da sich der 2. Formant vor allem bei unbetonten Silben und in Funktionswörtern in die gegensätzliche Richtung bewegt. Auffällig sind die deutlichen Unterschiede für [ɐ], weil dieser ein unbetonter und wenig Informationswert tragender Vokal ist, stark in seinen Formantfrequenzen streut und durchschnittlich nicht besonders lang ist.

Auch nicht-signifikante Veränderungen zeigen ein systematisches Verhalten, sodass ein gemeinsames Maß für F_1 und F_2 vermutlich zu deutlicheren Veränderungen geführt hätte. Auf ein solches Maß wurde jedoch bewusst verzichtet, da dann eine Normalisierung für den Vokalraum jeden Sprechers hätte durchgeführt werden müssen, um beide Formanten gleichwertig zu berücksichtigen. Auf die hier verwendete recht simple Weise sind die Ergebnisse einfacher nachzuvollziehen und robuster. Es wird deutlich, dass die meisten Vokal-Klassen einen dominanten Formanten besitzen, der sich dadurch auszeichnet, dass er sich signifikant, besonders stark und/oder in mehreren Bedingungen mit dem Tempo ändert, wie etwa F_1 für [aː] und F_2 für [eː].

Im Vergleich zur Segmentdauer, die klassischerweise als lokales Tempomaß angenommen wird, liegen die Resultate der Analysen im selben Bereich, weshalb

[5] In den Abbildungen als „O“ dargestellt (SAMPA).

erstere hier auch nicht dargestellt werden. Dies betrifft die signifikanten Ergebnisse der Hypothesentests und auch die R^2-Werte der Regressionen, wobei die PLSR häufiger zu leicht höheren Werten führt. Aufgrund der z-Transformation liegt die Varianz, die durch PLSR erklärt wird, im Bereich von nur etwa 10%. Ohne diese Transformation und mit Einbeziehung der Sprecheridentität steigt der R^2-Wert dann auf 0,4–0,6. Insofern sind lokales Tempo und individuelle, vor allem Geschlechterunterschiede bedeutende Faktoren für Formantfrequenzen. Jedoch wird aufgrund der großen Streuung über die PLSR keine zufriedenstellende Normierung der Formantwerte erreicht. Die teilweise hoch-signifikanten Ergebnisse beschreiben vielmehr den starken Zusammenhang zwischen Tempo und der Veränderung durchschnittlicher spektraler Lage der einzelnen Monophthonge. Im Folgenden wird untersucht, ob es auch einen Zusammenhang von sprecherspezifischem Tempo und Vokalreduktionen gibt.

9.2 Akustische Variation bei schnellen gegenüber langsamen Sprechern

Um den Einfluss von sprecherspezifischem Tempo auf Formantfrequenzen überprüfen zu können, muss der Anteil der Tempovariabilität innerhalb eines Sprechers minimiert werden. Eine Mittelung für jeden Sprecher würde nur zu maximal 10 Werten pro Bedingung führen. Da die Folge nicht-parametrische Tests wären, werden stattdessen nur Fälle verwendet, die pro Sprecher und Bedingung relativ durchschnittliche Tempi aufweisen ($\pm$ 1 SD um den jeweiligen Mittelwert). Die Unterschiede in der lokalen Sprechgeschwindigkeit zwischen den Sprechern sind nicht so groß sind wie die Variation innerhalb einer Person. Deshalb wird die Gruppe schneller Sprecher mit langsamen verglichen und dabei jede Person als randomisierter Faktor modelliert. Die gerade als hochsignifikanter Faktor ermittelte relative Sprechgeschwindigkeit bildet zur Kontrolle die Kovariate. Um Formantunterschiede im Geschlecht zu berücksichtigen, wurde diese binäre Variable mit einbezogen. Das Signifikanz-Niveau wird hier auf 1% festgelegt.

Für die Gruppen der schnellen und langsamen Sprecher ergeben sich mit neun signifikanten Ergebnissen weit weniger als für die relativen Variationen innerhalb der Personen (für die genauen Werte, siehe Anhang A.2). Dies mag auch daran liegen, dass die Tempodifferenz zwischen den beiden Gruppen etwa halb so groß ist, wie die zwischen schnellen und langsamen Phonrealisationen einer

Person.

Die Effekte betreffen im Gegensatz zu Kapitel 9.1 in nur zwei Fällen betonte Vokale in Inhaltswörtern. Nur bei [ə] sind beide Formanten für das Tempo signifikant verschieden. Für einen Fall ergibt sich eine Interaktion von Geschlecht und Sprechergruppe: Während die Lage der Formantfrequenzen für die Geschlechter natürlich auseinander fallen und keine weiteren Auswirkungen auf die Tempounterschiede haben, bedeutet der Nebeneffekt für [aː] in Funktionswörtern, dass sich F_1 nur innerhalb der weiblichen Sprecher unterscheidet. Bei fast allen Ergebnissen handelt es sich um Zentralisierungen. Ausnahmen bilden, wie für die relativen Unterschiede, F_1 für [ə] in Funktionswörtern. Das gilt auch für F_2 in Inhaltswörtern. Allerdings schwankt Schwa um das hypothetische Zentrum, sodass weder von De- noch Zentralisierung gesprochen werden kann.

9.3 Zentralisierung gegenüber verstärkter Koartikulation

Nachdem tempobedingte Vokalzentralisierungen für zahlreiche Vokal-Klassen nachgewiesen wurden, bleibt die Frage, ob es sich bei diesem Phänomen um tatsächliche Reduktion zum akustischen Zentrum handelt, oder ob es bloß ein Nebeneffekt der zahlreichen verschiedenen Kontexte darstellt. Eine Alternative bildet eine verstärkte Koartikulation, wie schon von Lindblom (1963) beschrieben. Die Überprüfung wird durch Kontrolle des lautlichen Kontextes durchgeführt. Dazu bietet sich F_2 an, der als besonders von lingualer Koartikulation beeinflusst angesehen wird. Er wird noch einmal bei betonten Realisationen von [ɔ] überprüft, die ja eine Denzentralisierung zeigten, die so erklärt werden soll.

Hier liegt F_2 im Mittel bei ca. 1000 Hz. Sollten von diesem schon recht extremen Bereich signifikante tempobedingte Veränderungen auftreten, die je nach Kontext gegensätzlich sind, wäre dies ein deutliches Zeichen für verstärkte Koartikulation. Dazu werden zwei gegensätzliche Kontexte verglichen, alveolare Konsonanten (mit Loki für Plosive von etwa 1800 Hz),[6] und alle Kombinationen ohne alveolare Umgebung. Da rein labiale Kontexte (700 Hz) nicht auftreten, werden diese und velare Konsonanten (1000 Hz für hintere Vokale) zusammen ausgewählt.

[6] Die Vergleichswerte stammen aus (Delattre et al., 1955).

Eine erneute z-Transformation für die Sprecher ist hier mit 1 – 6 Werten pro Person nicht möglich. Stattdessen werden getrennt für die Geschlechter einseitige Mann-Whitney Tests durchgeführt. Unabhängige Variable ist das lokale Tempo **schnell** gegenüber **langsam** mit dem Mittelwert als Gruppengrenze. Abhängige Variable bildet F_2 in Bark. Signifikanzniveau für diesen nicht-parametrischen Test ist wegen der geringen Fallzahlen und ungünstigen Kontexte 5%.

Das aus Kapitel 9.1 bekannte Verhalten eines Abfallens von F_2 wird für nicht alveolare Kontexte bestätigt (für Männer: $U(98, 124) = 7021$; $p < .05$; für Frauen: $U(69, 84) = 3473$; $p < .05$). Ein gegenteiliger Effekt tritt für alveolare Umgebungen auf. Hier steigt F_2 für Frauen an ($U(66, 40) = 974$; $p < .05$). Für Männer ist das Ansteigen von F_2 nicht signifikant ($U(53, 49) = 1098$; $p = .09$).

Für eine zweite Überprüfung bietet sich der Vergleich mit Gendrot und Adda-Decker (2005) an, die bei deutschem Material keine Koartikulation für [a] nachweisen konnten. Für unbetonte [a] wurden Funktionswörter gewählt, da hier eine größere Auswahl an Fällen vorliegt. Im Mittel liegen die F_2-Werte knapp unter dem theoretischen Zentrum vom 1500 Hz für Männer und 1650 Hz für Frauen. Als Umgebung für Zentralisierungen wurden alveolare Konsonanten ausgewählt. Für eine erwartete Dezentralisierung als Zeichen von Koartikulation sind keine reinen labialen Umgebungen vorhanden. Deswegen werden labial-[a]-alveolar Verbindungen untersucht. Auch hier werden wieder Mann-Whitney Tests durchgeführt.

F_2 verändert sich im Gegensatz zu Gendrot und Adda-Decker (2005) signifikant für den alveolaren Kontext: Die Werte für [a] in dieser Bedingung sind für die Gruppe schneller geäußerter Phone höher (für Männer: $U(405, 406) = 69116$; $p < .001$; für Frauen: $U(301, 294) = 40632$; $p < .05$), verändern sich aber für labiale prä-vokalische Konsonanten nicht (für Männer: $U(35, 39) = 646$; $p = .65$; für Frauen: $U(34, 23) = 358$; $p = .71$).

9.4 Zusammenfassung und Diskussion

Die natürlich auftretenden intrapersonellen Schwankungen in lokaler Sprechgeschwindigkeit sind sehr groß und treten bei vielen Vokalen zusammen mit hoch-signifikanten relativen spektralen Veränderungen auf. Diese Veränderungen zeigen sich im Mittel als Zentralisierungen. Eine genaue Berechnung des akustischen Vokalraumes für schnelle gegenüber langsamen Phonen entfällt: Zu

deutlich sind die Zentralisierungen der Eckvokale, um eine Verkleinerung des Vokalraums nachmessen zu müssen.

Besonders betroffen sind betonte Vokale in Inhaltswörtern, die mit ihrer extremen Lage auch mehr Raum für Veränderungen aufweisen, allerdings auch bedeutsamer für die Worterkennung sind. [eː], [aː], [a], [ɔ] sind die Vokal-Klassen, die diese Veränderungen in allen drei Bedingungen (betonte und unbetonte Inhaltswörter, Funktionswörter) zeigen. Zusammen mit [ɐ] bilden [aː] und [a] die Gruppe von Monophthongen mit zentraler Zungenlage und (halb-)tiefer Zungenhöhe, die als einzige komplett und in jeder überprüften Bedingung diese Zentralisierungen aufweist. F_1 ist insgesamt öfter von Temposchwankungen betroffen als F_2. Da Vokale mit tiefer Zungenhöhe mehr Raum für F_1 Zentralisierungen bieten, ergibt sich hier eine Begründung für den robusten Effekt bei diesen Vokalen.

Die Ausnahme von der Tendenz zum akustischen Zentrum bildet [ɔ] in betonten Inhaltswörtern für F_2. In Funktionswörtern zeigt sich aber wieder eine Zentralisierung. Dieser Effekt kann nach der Überprüfung mit verschiedenen konsonantischen Umgebungen als Ausnahme von der Zentralisierung gewertet werden. Da die Ursache der spektralen Veränderungen verstärkte Koartikulation darstellt, setzen sich in betonten Silben im Durchschnitt Kontexte durch, die F_2 verringern. Ähnliches muss für Schwa gelten, dessen F_1-Werte für hohe Tempi sinkt.

Den relativen Temposchwankungen stehen die Unterschiede zwischen schnellen und langsamen Sprechern gegenüber, die in weniger Bedingungen signifikant sind. Hier zeigt sich eine deutliche Verteilung von tempobedingten Veränderungen von F_1 für tiefe und von F_2 für vordere Vokale. Es werden also gerade die Formanten von schnellen Sprechern reduziert, die für die jeweiligen Vokale besonders stark vom Zentrum abweichen. Bei den Sprecherunterschieden ergibt sich auch ein Effekt für [ɪ] bei unbetonten Silben, der als einziger nicht für intrapersonelle Unterschiede (Kapitel 9.1) auftritt.

Im Vergleich zu anderen akustischen Analysen (vgl. Kapitel 3.1.1) sind die vorliegenden Ergebnisse eine Bestätigung ältester (Lindblom, 1963) und neuester Studien (Gendrot und Adda-Decker, 2005): Durch die Kontrolle des Sprechstils und getrennte Analysen für Betonung und Wortart lassen sich die Formantveränderungen klar der Tempovariation zuordnen. Dies ist ein natürliches Verhalten für die provozierte Kommunikationssituation, da weder die intra-, noch interper-

sonellen Temposchwankungen von außen vorgegeben wurden. Intrapersonelle – und in einem geringeren Rahmen auch interpersonelle – Tempoerhöhung führt zu spektralen Reduktionen. Bei den beiden zentralen Vokalen [ə], [ɐ] wird F_1 auch über das Zentrum hinaus verringert. Ansonsten kann aber durchschnittlich von einer Zentralisierung gesprochen werden. Unter Berücksichtigung von Ergebnissen anderer Untersuchungen zeigt sich, dass es sich bei diesen Tempoeffekten allerdings um verstärkte Koartikulation mit der konsonantischen Umgebung handelt. Es ist unwahrscheinlich, dass diese Effekte sprachspezifisch sind, da außer dem Deutschen auch das Französische (Gendrot und Adda-Decker, 2005) und Schwedische (Lindblom, 1963) betroffen sind. Vielmehr zeigt es sich für spontane authentische Sprache, während es sich bei Stack et al. (2006), der keinen solchen Effekt findet, um ein Experiment mit Laborsprache handelt.

Die aus der Literatur benannten Unterschiede zwischen intra- und interpersonellen Tempovariationen (Turner et al., 1995; Tsao et al., 2006), also keine signifikanten spektralen Reduktionen für schnelle gegenüber langsamen Sprechern bei gelesener Sprache, können hier in ihrer extremen Ausprägung nicht nachvollzogen werden. Laut Tsao et al. (2006) zeigen die langsamen Sprecher ein etwa so hohes maximales Tempo wie die schnellen im Mittel. Die für die vorliegende Arbeit verwendeten Gruppen weichen trotz der Begrenzung auf eine Standardabweichung um den jeweiligen Mittelwert nicht so stark voneinander ab, weisen aber dennoch signifikante Unterschiede auf. Möglicherweise handelt es sich um einen Gegensatz von Lese- zu Spontansprache.

Es wird aber deutlich, dass der Effekt tempobedingter Reduktionen für intrapersonelle Schwankungen weit mehr Bedingungen betrifft als interpersonelle. Die Ergebnisse für relative Veränderungen gelten für die Gruppe aller Sprecher und lassen sich aufgrund des Zufallsfaktors verallgemeinern. Dagegen betreffen die interpersonellen Ergebnisse erst einmal nur die Gruppen schneller und langsamer Sprecher bei individuell mittlerem Tempo. Ein Vergleich beider Effekte ist schwierig. Weil die zu berücksichtigenden intrapersonellen Tempounterschiede etwa doppelt so groß sind wie solche für interpersonelle ($2 \cdot SD = 1{,}35$ gegenüber der Differenz $langsam - schnell = 1{,}3$), muss hier eine Aussage zur Gewichtung der signifikanten Ergebnisse entfallen. Wichtig ist das Ergebnis, dass in einigen Bedingungen auch Sprecherunterschiede signifikant sind. Hauptsächlich lassen sich aber tempobedingte Veränderungen der ersten beiden Formanten über relative Tempovariation erklären. Die hier verwendeten Daten haben im Vorfeld ohne Trennung zwischen intra- und interpersonelle Unterschiede (Weiss

(2005b), absolute Formantwerte in Bark, absolutes Tempo in $^{Sil'}/_{s}$) zu fast identischen Effekten wie die relativen Veränderungen hier geführt. Die sich ergebende Dominanz für spektrale Reduktionen durch intrapersonelle Tempoerhöhung gilt allerdings nur für das hier untersuchte Material. In anderen Kommunikationssituationen mag es stärkere individuelle Unterschiede im mittleren Tempo geben, die dann auch stärkere Folgen für die spektralen Eigenschaften der Monophthonge haben.

Als Folge für die Wahrnehmung ergibt sich die besondere Bedeutung des lokalen Charakters von Sprechgeschwindigkeit, durch den sich erst intrapersonelle Unterschiede ergeben. Die hier gezeigten tempobedingten spektralen Variationen lassen sich nicht über individuelle Unterschiede erklären. Lokale Sprechgeschwindigkeit zeigt sich dagegen als eigenständiger Faktor, der mit spektraler Vokalvariation zusammenhängt.

Zwar ist nicht geklärt, ob bei der Verarbeitung solcher Variationen tatsächlich Informationen, wie sie PLSR darstellt, genutzt werden. Tempovariation führt jedoch zu schlechterer und langsamerer Worterkennung. Die Unterschiede von 80 Hz bis 150 Hz bezüglich der Formantfrequenz zwischen mittleren schnellen und mittleren langsamen Fällen sind zwar nicht sehr groß, aber in einem Bereich, der nicht nur statistisch relevant ist, sondern bereits zu hörbaren Qualitätsunterschieden führt. In neueren Ansätzen (vgl. Kapitel 4.2) wird jedenfalls davon ausgegangen, dass praktisch alle zur Verfügung stehenden Informationen bei der Verarbeitung verwendet werden. Es gibt keinen Grund anzunehmen, dass dies für die hier gezeigten Ergebnisse anders sein sollte. Dies gilt insbesondere, da das Register durch die Kommunikationssituation, in der das Korpus produziert wurde, für Spontansprache schon relativ gehoben ist; also weit informellere und reduziertere Sprache im Alltag auftritt. Dagegen streuen die Daten auch mit Einbeziehung von lokalem Tempo in einem Maße, dass sich keine invarianten Parameter ergeben.

Phondauern führen bei den hier verwendeten Daten nicht zu besseren Ergebnissen als die PLSR. Sollte eine perzeptive Relevanz der hier vorgestellten akustischen Effekte vorliegen, stellt sich die Frage, wie der Zusammenhang von lokalem Tempo und spektralen Vokalinformationen verarbeitet würde. Im Vergleich zu Phondauern bedarf die PLSR eines größeres Zeitfensters von zumindest der Silbe, um zu entstehen. Deshalb müsste diese Art von lokaler Tempoinformation separat von spektralen Informationen, also extrinsisch verarbeitet werden (vgl. Kapitel 4.1). Für Phondauern ist dies nicht zwingend notwendig. Die Korrelati-

on zwischen Vokaltarget und -dauer ist auch über ein einziges Maß, nämlich die Steigungen der Formanttransitionen erfassbar, was dann intrinsisch verarbeitet werden würde.

10 Spektrale Analyse stimmloser Frikative

Nach den Vokalen werden nun ausgewählte Konsonanten auf spektrale Veränderungen durch lokales Tempo untersucht. Hierbei wird ein Parameter extrahiert, die *spektrale Balance* oder auch *center of gravity* (COG) genannt (vgl. Kapitel 3.1.3). Das COG ist wie folgt definiert:

$$COG = \frac{\int f \cdot E_f \cdot df}{\int E_f \cdot df}$$

Hierbei entspricht f der Frequenz und E_f der Energie in dieser Frequenz.

Bei stimmlosen Frikativen wurden für die Dauer von 30 – 70% der Segmentlänge alle 5 ms Schmalband-Spektren mit der Fast-Fourier-Transformation berechnet. Diese wurden gemittelt und daraus die spektrale Balance (in Hz) erhoben. Im Gegensatz zu den Vokalformanten sind diese Werte nicht direkt mit Daten anderer Untersuchungen vergleichbar, da sie vom genutzten Freqenzbereich (hier 0 – 8 kHz) und der Art der Energieerhebung abhängen.

Die Auswertung erfolgt getrennt für sechs Konsonanten-Klassen, Betonung und Wortart. Da [χ] nicht annotiert ist, wurde die Trennung gegenüber [x] über den prä-konsonantischen Vokal durchgeführt, wie in Kohler (1995) beschrieben. Die abhängige Variable bildet das COG und wird in jeder Bedingung über ihren jeweiligen Mittelwert pro Sprecher normiert. Ausgeschlossen werden alle [h], da glottale Frikative im Deutschen zumeist als behauchte Qualität des nachfolgenden Vokals auftreten (vgl. dazu auch Ladefoged und Maddieson, 1996). Als Kontrollvariable wird das **Geschlecht** einbezogen, sowie die Unterscheidung nach **Auftreten** des Frikativs in Onset oder Koda. Es werden 18472 Fälle untersucht.

Unterschiede zwischen den einzelnen Sprechern sind nach der Normierung unbedeutend. Das Geschlecht zeigt sich in zwei Fällen signifikant: Bei betonten [ç] weisen weibliche Sprecher eine stärkere Verringerung im COG (etwa -68 Hz)[1]

[1] Wie auch bei den Vokalen sind diese absoluten Werte über nicht normierte Daten berechnet.

für ansteigendes Tempo auf als männliche (-31 Hz; $F(1,708) = 12.75$; $p < .001$). Ein vergleichbarer Effekt tritt für betonte [χ] auf, wo ebenfalls der Unterschied zwischen weiblichen (-46 Hz) und männlichen Sprechern (-20 Hz) signifikant ist ($F(1,315) = 4.29$; $p < .05$).

Für alle untersuchten Artikulationsorte zeigen sich signifikante Verminderungen des COG mit ansteigendem Tempo. Für betonte [ʃ] und [x] sind sie allerdings nur auf dem 5%-Niveau signifikant (siehe Anhang A.3). Letztere bilden mit 69 Fällen die kleinste der untersuchten Gruppen. Die nicht signifikanten Analysen betreffen ausschließlich [ʃ]. Dieser Frikativ zeigt nur in der Bedingung mit den meisten Fällen (523, betonte Silben) den gerade beschriebenen Effekt. Die genauen Werte für alle signifikanten Ergebnisse befinden sich im Anhang A.3. Statistiken mit der Segmentdauer anstatt PLSR erbringen vergleichbare Ergebnisse mit geringeren F-Werten. Einziger nennenswerter Unterschied ist das Verfehlen eines signifikanten Einflusses der Segmentdauer auf das COG für [ʃ].

In Abbildung 10.1 sind die Mittelwerte des COG für die jeweils langsam und schnell geäußerten Frikative aufgetragen; getrennt für die Bedingungen **Inhaltswörter betont** (bet.) und **unbetont** (unbt.), sowie **Funktionswörter** (Funk). Die Mittelwerte repräsentieren die Fälle mit den jeweils 30% höchsten und niedrigsten Werten in der PLSR. Diese Auswahl stellt auch die Ursache für die geringen Mittelwertunterschiede für die Bedingung „f Funk“ dar. Somit dient diese Abbildung der Illustration, während im Anhang A.3 sinnvollere Erwartungswerte aufgrund aller Daten und der statistischen Analyse zu finden sind. Nur Bedingungen mit einem signifikanten Ergebnis sind hier abgebildet.

Mit diesen relativen Maßen werden sprecherübergreifende Reduktionen mit ansteigendem lokalen Tempo bewiesen. Sie liegen selbst für maximale Tempovariation eines Sprechers von 5,4 $^{\text{Sil}'}/_{\text{s}}$ (bei $\pm$ 2-facher Standardabweichung) im Mittel bei etwa 100 Hz. Dagegen beträgt die maximale Differenz zwischen den Sprechern im COG eines Konsonanten etwa 590 Hz (s), 660 Hz (ʃ), 720 Hz (ç), 430 Hz (f), 375 Hz (x) und 323 Hz (χ). Bei über dem 4-fachen von Sprecher- gegenüber Tempo-Unterschieden (mit Ausnahme von [x]) wird deutlich, dass die hier dargestellten Effekte gegenüber sprecherspezifischen Unterschieden nicht ins Gewicht fallen. Dies ist vielleicht ein Grund für die spärliche Anzahl solcher Untersuchungen in der Literatur. Dagegen wurde bereits vielfach auf individuelle Unterschiede hingewiesen (vgl. Shadle, 1990).

Ursache für diese relative Stabilität der spektralen Information gegenüber Voka-

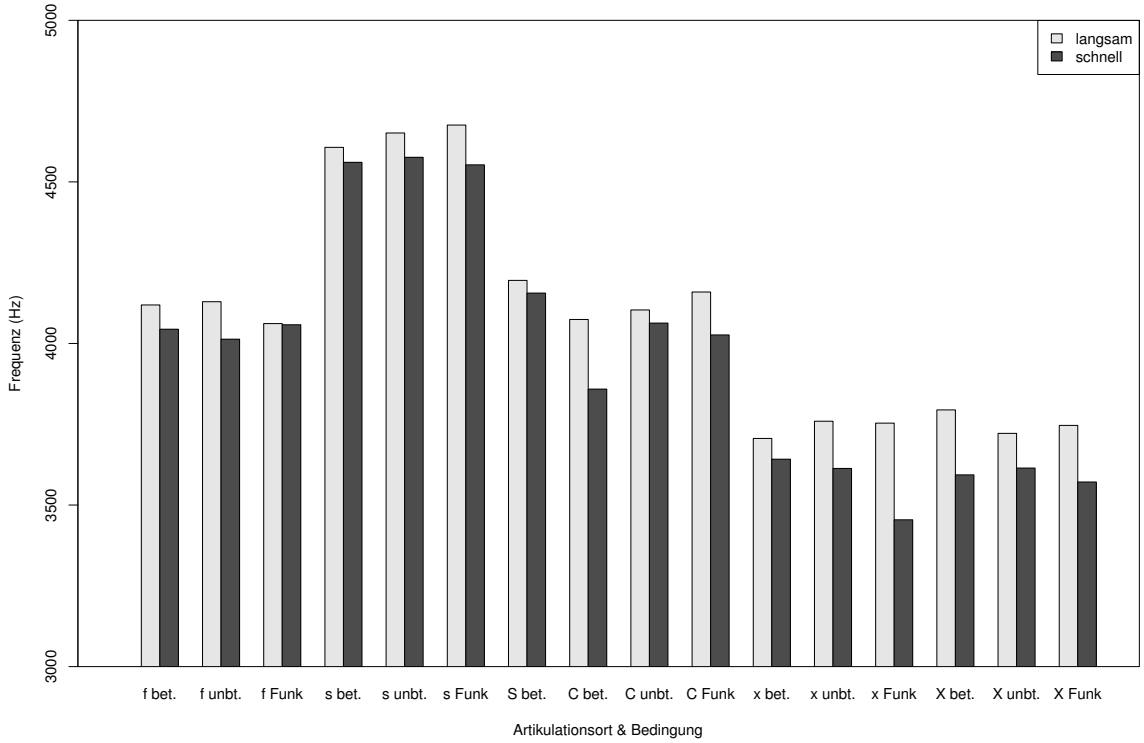

Abbildung 10.1: Mittleres COG für alle signifikanten Bedingungen

len liegt wohl in den genau definierten Artikulationsorten von Frikativen (Ladefoged und Maddieson, 1996), sodass sie robuster auf Tempovariation reagieren als andere Lautklassen, während artikulatorische Unterschiede in Produktion und Mundraum in spezifischen Spektren für einzelne Sprecher resultieren. Diese relative Stabilität gilt jedoch nur für Frikative, die auch in der Transkription als Frikativ annotiert sind. Schließlich sind Konsonanten in der Silbenkoda von Reduktionen und Elisionen stärker betroffen als Vokale (vgl. Kapitel 3.2). Reduktionen, die zu einer abweichenden Annotation führen, wie tempobedingte Änderungen der Artikulationsart, werden in diesem Kapitel gerade nicht erfasst.

Für eine Phonemidentifikation werden die hier dargestellten Effekte allerdings keine Rolle spielen, da sich im Gegensatz zu Formanten bei Monophthongen die Verteilungen des COG für verschiedene Phoneme kaum überschneiden. Einzig [ʃ] und [f] zeigen ähnliche Werte. Diese beiden Artikulationsorte sind aber anhand ihrer spektralen Struktur deutlich zu unterscheiden.

Die binäre Kontrollvariable zur Silbenposition in Onset oder Koda zeigt in keinem der analysierten Fälle einen signifikanten Einfluss. Dieses Ergebnis bedeutet, dass sich der beobachtete Zusammenhang zwischen lokalem Tempo und COG nicht über diese Position erklären lässt, da Frikative im Onset nicht ge-

nerell weniger von diesem Effekt betroffen sind als solche in der Koda. Dieses Ergebnis ist insofern bemerkenswert, weil zum einen dieser Unterschied für symbolisch erfasste Reduktionen von Konsonanten relevant ist. Zum anderen scheidet damit ein möglicher Einflussfaktor auf Tempo und Spektrum aus, der im Sinne von van Son und van Santen (2005) Ursache konsonantischer Reduktion inklusive Dauerverkürzung sein könnte.

11 Analyse von Wortrealisierungen anhand der Transkription

In diesem Kapitel werden Aussprachevariationen, die sich über die symbolische Umschrift erfassen lassen, auf Abhängigkeit zum lokalen Tempo untersucht. Einer Untersuchung aller Wortformen auf Zusammenhänge zwischen Aussprache und lokalem Tempo würde kaum zu aussagekräftigen Ergebnissen führen, da in diesem Fall nicht zwischen intrinsischem Tempo, phonologischen Eigenschaften der Wörter und sprachlicher Realisierung unterschieden werden könnte. Für Ergebnisse solcher generellen Analysen vergleiche stattdessen Kapitel 3.2. Um den Zusammenhang von PLSR und Aussprachevariation getrennt für verschiedene Wortformen zu überprüfen, bedarf es zahlreicher Realisierungen eines Wortes. Für die 34 häufigsten Wortformen (solche mit über 200 Realisierungen im Korpus, 15338 insgesamt, siehe Appendix A.4) wurde vermerkt, wie viele Reduktionen,[1] Elisionen und Epenthesen sich in jeder Realisierung wiederfinden und mit dem jeweiligen gemittelten Tempo (gemessen im Bereich 20 – 80% der Wortlänge) in Beziehung gesetzt. Dafür wurden klitische Formen von der gemeinsamen Analyse ausgeschlossen und stattdessen getrennt behandelt. Bei den meisten Wörtern handelt es sich um Einsilber, nur sieben bestehen aus zwei Silben, sodass 98% der Realisierungen kürzer als 500 ms sind. Insofern bleibt das Tempo trotz Mittelung der Werte lokal. Da sich nur zwei Inhaltswörter („machen“ und „gut“) in dieser Gruppe befinden, werden die Analysen nicht getrennt für die Wortart durchgeführt. Nur vollständige Wörter ohne Häsitationen oder Verbesserungen werden hier untersucht.

[1] Streng genommen bezeichnet eine phonetische Reduktion auf Symbolebene eine segmentale Abschwächung, wie etwa eines Vollvokals zu einem Schwa oder eines Plosivs zu einem Frikativ oder Approximanten. Im vorliegenden Kapitel werden verschiedene phonetische Prozesse, die zu einer Änderung in der Annotation führen, also Assimilationen und Reduktionen, nicht von einander unterschieden und nur im Kontrast zu Elisionen, also dem Wegfall von Symbolen, als Reduktionen bezeichnet, so wie sie auch im Korpus annotiert sind.

11.1 Abweichen von kanonischen Wortrealisierungen

Wie bei den spektralen Analysen, wo keine besonderen geschlechtsbedingten Unterschiede auftreten, soll auch in diesem Kapitel überprüft werden, ob sich Männer und Frauen bei ihren Wortrealisierungen unterscheiden. Für die Kovarianzanalyse wird **Geschlecht** also als Kontrollvariable behandelt, und die einzelnen Wortformen werden als randomisierter Faktor innerhalb der Sprecher hierarchisch modelliert.

Die Ergebnisse der Statistik zeigen hochsignifikante Einflüsse von Tempo auf Elisionen[2] ($F(1, 14271) = 988.10$; $p < .001$) sowie auch auf Reduktionen ($F(1, 14271) = 329.14$; $p < .001$). Beide Prozesse treten bei steigendem Tempo häufiger in einer Realisierung auf. Bei Elisionen bedeutet dies vor allem das Auftreten von genau einer Elision. Kein signifikanter Effekt ergibt sich für den Zusammenhang zwischen Epenthesen und Tempo ($F(1, 14271) = 1.12$; $p = .30$). Das Auftreten von Epenthesen ist vom jeweiligen Wort abhängig ($F(1, 33) = 6.74$; $p < .001$).[3] 28 der 32 Sprecher haben Epenthesen produziert, insgesamt aber nur 58 Fälle bei insgesamt 19 verschiedenen Wörtern. Unterschiede im Geschlecht zeigen sich nur für Elisionen ($F(1, 30) = 7.44$; $p < .05$), die für weibliche Sprecher seltener sind, ohne den signifikanten Effekt vermehrter Elisionen bei steigendem Tempo zu beeinflussen.

Wortdauern modellieren Reduktionen nicht so gut wie PLSR. Was den Zusammenhang mit der Anzahl von Reduktionen betrifft, zeigt ein *Likelihood Ratio Test*, dass sich die beiden Variablen **Wortdauer** oder **PLSR** signifikant voneinander unterscheiden ($\chi^2(1) = 769.67$; $p < .001$), wobei die PLSR den Zusammenhang zu Reduktionen besser modelliert als die Wortdauern.[4]

Für eine bessere Quantifizierung der Aussprachevariationen als über bloße Häufigkeiten wird ein linguistisches Maß für die Abweichung einer Wortrealisierung zur kanonischen Aussprache verwendet. Der errechnete Wert hängt von den jeweiligen Unterschieden in Form von phonetischen Beschreibungsmerkmalen ab.

[2] In diesem Fall wird eine Elision als fehlendes Symbol auf der Ebene der phonetischen Umschrift gegenüber der kanonischen angesehen. Da im Kielkorpus der Glottalplosiv auf der kanonischen Ebene obligatorisch vor ansonsten silbeninitiale Vokale gesetzt wird, gilt sein Fehlen als Elision.

[3] **Wort** wurde hierfür als normale Variable, nicht als Zufallsfaktor betrachtet.

[4] Beide Modelle auf Signifikanz miteinander zu vergleichen, ist mit Standardmethoden nicht möglich.

So wird z. B. eine Schwa-Tilgung als weniger starke Veränderung angesehen als die eines vollen Vokals (Herrgen und Schmidt, 1989). Der genaue Wert für ein Wort ergibt sich aus der Addition der Phon-Unterschiede und berücksichtigt sowohl Reduktionen als auch Elisionen. Dieses quantitative Maß zeigt sich valide und reliabel im Bereich der Dialektforschung und stellt eine Adaption von Vieregge et al. (1984) dar, die ein vergleichbares Verfahren zur Einschätzung von Abweichungen unter Transkribenden verwenden. Für die Berechnung dieses Maßes verwenden Herrgen und Schmidt (1989) die Transkriptionen von Krech et al. (1982) als kanonische Aussprache, in deren Wörterbuch ein Ausfall des Glottalplosivs in unbetonte Silben fließender Rede – also für fast alle hier betroffenen Fälle – nicht als Abweichung erfasst wird (vgl. Krech et al. (1982), S. 74).

Abweichungen von der kanonischen Form treten schon bei niedrigen Tempi auf, was für diese kurzen und häufigen Wörter nicht ungewöhnlich ist (vgl. Kapitel 3.2). Bei der Verwendung des Maßes nach Herrgen und Schmidt (1989) zeigt sich, dass ein verstärktes Abweichen von der kanonischen Form signifikant für die Haupteffekte **Tempo** ($F(1, 14271) = 735.50$; $p < .001$) und **Wort** ($F(33, 1007) = 58.7$; $p < .001$) ist. Während die Distanz von einer kanonischen Aussprache mit steigendem Tempo generell zunimmt, ist diese Zunahme für weibliche Sprecher insgesamt geringer ($F(1, 14271) = 9.8$; $p < .01$; für die Interaktion **Tempo** und **Geschlecht**). Dieses Zusammenwirken von **Geschlecht** und **Tempo** lässt sich über die Zusammenlegung von Elisionen und Reduktionen zu dem neuen Maß erklären, da für Elisionen ein Effekt für das Geschlecht ja bereits ermittelt wurde. Die fehlende Berücksichtigung von Glottalelisionen verändert nicht den signifikanten Zusammenhang von Tempo und Abweichen von der kanonischen Form.

Die Unterschiede zwischen den einzelnen Wörtern sind groß. Besonders Realisierungen von „ein“, „vielleicht“, „ist“, „Ihnen“, „machen“, „nicht“, „und“ sowie „wäre“ fallen allgemein durch höhere Werte und damit stärkere Abweichungen zur kanonischen Aussprache auf. Der Nebeneffekt zwischen **Tempo** und **Wort** ($F(1, 14271) = 13.3$; $p < .001$) zeigt, dass die Stärke des Tempoeinflusses signifikant von der Identität des Wortes abhängt.

11.2 Tempo und Auftreten klitischer Formen

Klitische Formen von häufigen Wörtern (nicht das lexikalisierte „am“) werden in den bisher beschriebenen Analysen nicht berücksichtigt. Vergleicht man die-

se 103 Realisierungen bei 10 Wortformen mit den anderen Realisierungen im Bezug auf ihr Tempo, wird deutlich, dass Wortrealisierungen mit klitischen Zusätzen 0,77 $^{\text{Sil}'}/_{\text{s}}$ langsamer sind als solche derselben Wortform ohne Zusätze ($F(1, 4431) = 23.4$; $p < .001$). Dieses Ergebnis ist fast selbstverständlich, da ja den Wörtern in allen Fällen ein [s] angehängt wurde. Hier zeigt sich aber auch, dass von diesem Effekt nur Wörter betroffen sind, die generell zu den schneller artikulierten der 34 gehören ($F(1, 14367) = 215.9$; $p < .001$). Allerdings darf dieses Ergebnis nicht als Hinweis interpretiert werden, Wörter, die ein höheres lokales Tempo aufweisen, würden deshalb eher mit klitischen Formen des Nachfolgewortes realisiert. Hier stellt die phonologische Beschaffenheit der Wörter eine Ursache für die Tempounterschiede dar; schließlich ist das Auftreten von klitischen Formen auf bestimmte Wörter und syntaktische Positionen wie etwa die Wortkombinationen „wie es“ begrenzt. Die Einbeziehung dieser Formen in die bereits durchgeführten Statistiken verändert die signifikanten Tempoeffekte im vorherigen Kapitel nicht.

11.3 Auftreten von Nasalierungen und Laryngalisierungen

In der phonetischen Umschrift sind auch Nasalierungen und Laryngalisierungen vermerkt. Dazu ist anzumerken, dass im Kielkorpus Nasalierungen nur bei Vokalen annotiert sind, falls gleichzeitig auch ein benachbarter Nasal getilgt ist. Dies gilt allerdings nicht für Laryngalisierungen. Für die Wortformen, deren Transkription das Merkmal Laryngalisierung oder Nasalierung aufweisen, ist auch ihre Sprechgeschwindigkeit signifikant abweichend ($F(2, 10846) = 139.9$; $p < .001$).

Während ein Fall, der eine Nasalierung enthält, durchschnittlich 1,3 $^{\text{Sil}'}/_{\text{s}}$ schneller ist als eine Wortrealisierung ohne ($F(1, 2654) = 34.6$; $p < .001$, mit **Wort** als Zufallsfaktor), ist für Laryngalisierungen das Gegenteil der Fall ($F(1, 5114) = 462, 9$; $p < .001$, -1,1 $^{\text{Sil}'}/_{\text{s}}$, mit **Wort** als Zufallsfaktor).[5]

Der Zusammenhang der Annotation von Nasalität bei einem Vokal bei stei-

[5]Wegen der geringen Fallzahlen wurden einseitige Mann-Whitney Tests zur Kontrolle durchgeführt, die die Ergebnisse der beiden F-Tests bestätigen. Bei den Laryngalisierungen wurden 18 Fälle ausgeschlossen, bei denen die Laryngalisierung nicht wortinitial auftrat und damit nicht direkt über das Auftreten oder die Substitution eines Glottalplosivs erklärt werden kann. Diese Fälle sind damit anderen Ursprungs, z. B. Substitutionen stimmloser Plosive (vgl. Kohler, 1984a), werden jedoch aufgrund der geringen Anzahl nicht untersucht.

gendem Tempo kann über die Elision des benachbarten Nasals erklärt werden. Nach Überprüfung der Transkriptionen von Fällen mit Larygalisierungen ergeben sich jedoch keine Auffälligkeiten in dem Auftreten dieser Annotation. Insbesondere ist die Laryngalisierung unabhängig vom Auftreten eines Glottalplosivs ($\chi^2(1) = 2.9$; $p = .09$), sodass sich die Nullhypothese: *„Laryngalisierung tritt unabhängig von Realisierungen des Glottalplosivs auf."*, nicht ablehnen lässt. Statistisch lässt sich damit auch ein Zusammenhang von Laryngalisierungen als Ersatz eines Glottalplosivs nicht nachweisen. Die Laryngalisierung des Vokals selbst könnte die Ursache für das langsamere Tempo darstellen. Das Ergebnis betreffend der Laryngalisierungen steht in Einklang damit, dass diese und nicht Glottalplosive die Norm für Silbenonsets vor Vokalen in fließender Rede darstellen (vgl. Kohler (1984a); Janker et al. (1999)).

11.4 Darstellung der Ergebnisse anhand einzelner Wörter

Während die Übersichtsanalyse in Kapitel 11.1 einen generellen Trend für eine stärkere Ausspracheabweichung bei ansteigendem Tempo zeigt, soll in den Einzelanalysen überprüft werden, bei welchen Wörtern dieser Effekt tatsächlich auftritt und um welche Aussprachevarianten es sich dabei handelt.[6] Im Gegensatz zum Maß von Herrgen und Schmidt (1989) wird im vorliegenden Kapitel die im Kielkorpus angegebene kanonische Aussprache als Referenz verwendet, was den wortinitialen prävokalischen Glottalverschluss beinhaltet.[7]

11.4.1 Wörter ohne tempobedingte Aussprachevariationen

Bei diesen Einzelanalysen zeigen insgesamt 6 der 34 Wörter keinen signifikanten Zusammenhang zwischen Ansteigen des Tempos und des Abweichens von einer kanonischen Aussprache (vgl. Tabelle A.4). Diese sechs Ausnahmen sollen nun

[6]Streng genommen stellt das Maß nach Herrgen und Schmidt (1989) keine intervallskalierte Variable dar, weil es auf positive Werte mit Stufen von 0,5 beschränkt ist, und sein Auftreten von den möglichen Realisierungen der jeweiligen Worte abhängt. Dennoch ergeben F-Tests sinnvolle Ergebnisse, wie Kontrollstatistiken gezeigt haben: Die Ergebnisse weichen in ihrer Signifikanz nicht von nicht-parametrischen Tests oder Statistiken mit Tempo als z-transformierte PLSR ab. Einen Vorteil von F-Tests stellt die Ermittelung eines Erwartungswertes dar, sowie die Verringerung der Analysen, da hier nicht die Daten auf Gruppen bezüglich Tempo und Abweichwert aufgeteilt werden müssen. Zudem wird hier der Faktor **Sprecher** berücksichtigt, was bei nicht-parametrischen Verfahren so nicht möglich ist.

[7]Unter den untersuchten Wortformen befindet sich keine mit silbeninitialen prävokalischen Glottalverschluss.

genauer untersucht werden.

Eine davon bildet das Wort „am“, das auch unabhängig vom Tempo kaum starke Veränderungen aufweist, was natürlich auch an der Kürze des Wortes liegt. Die Elision des Glottalverschlusses bildet hier die Regel, da sie über doppelt so häufig auftritt (291) wie die kanonische Form (127). Außer der Variante [m̩] (34) fallen keine anderen Realisierungen ins Gewicht. Diese Fälle mit silbischem Nasal sind zwar durchschnittlich schneller als die der kanonischen Aussprache, dieser Unterschied ist allerdings nicht signifikant ($F(1,98) = 1.26$; $p = .27$).[8] Dagegen weisen Fälle mit Elision des Glottalverschlusses, die ja in dem Maß von Herrgen und Schmidt (1989) nicht berücksichtigt wird, ein signifikant höheres Tempo als kanonische Realisierungen auf ($F(1,385) = 78.10$; $p < .001$). Dieses Ergebnis steht im Einklang mit dem allgemeinen Phänomen für diese häufigen Wörter, dass initiale Glottalplosive eher bei niedrigem Tempo auftreten (vgl. letztes Kapitel), obwohl sie nicht die tatsächliche Aussprachenorm darstellen (vgl. Kohler, 1984a). Insofern kann dieser Effekt als übergenaue Artikulation bei niedrigem Tempo angesehen werden. Wünschenswert wäre die Überprüfung eines breiteren Tempokontextes, in dem die „am“-Realisierungen eingebettet sind, um zu überprüfen, ob ein höheres Tempo im Kontext mit der Elision des Glottalverschlusses einhergeht. Da dieser jedoch schwer zu kontrollieren ist, bietet sich als Alternative an, die Segmentdauern von [a] und [m] als zusätzliches Tempomaß heranzuziehen. Es zeigt sich eine signifikante Verlängerung des Nasals bei Auftreten des Glottalplosivs ($F(1,96) = 31.51$, $p < .001$). Er ist mit 55 ms etwa ein drittel länger als in Wortrealisierungen ohne [ʔ]. Die Dauern des Vokals ändern sich dagegen nicht ($F(1,96) = 0.88$, $p = .38$).

Das Wort „wäre“ wird zusätzlich zur kanonischen Aussprache (130) auch als [ve͡ɐ] (73) realisiert. Beide Varianten unterscheiden sicht nicht bezüglich ihrer PLSR. Hier zeigt sich der Einfluss der Silbenlänge auf das lokale Tempo: Da die kurze zweite Silbe durch die /ʀ/-Tilgung wegfällt, ändert sich die mittlere PLSR nicht ($F(1,173) = 1.76$; $p = .19$), während die Wortdauer für die einsilbigen Realisierungen kürzer ist ($F(1,173) = 92.62$; $p < .001$).

Im Unterschied zu den gerade genannten Wörtern weist „vielleicht“ vier Varianten auf. Abgesehen von der kanonischen Form (85), wird häufig das [ɪ] elidiert (79), seltener der finale Plosiv (14) oder beide (20). Weitere Varianten sind nicht

[8] In dieser und folgenden zusätzlichen Statistiken mit wenigen relevanten Varianten wird die Abweichung von einer kanonischen Aussprache nicht wie in Anhang A.4 als Kovariate, sondern als Faktor behandelt, um Post-Hoc Vergleiche durchführen zu können.

nennenswert (7). Wie beim „wäre“ zeigt sich der Ursprung der PLSR in den Silben- und Phondauern. Mit der Tilgung des [ɪ] wird die jeweilige Aussprache oft einsilbig realisiert, sodass sich die Varianten nicht durch höheres Tempo auszeichnen können, da mögliche Segmentverkürzungen durch die verlängerte übrige Silbe neutralisiert werden.[9] Da sich die Wortdauer fast zwangsläufig mit Elisionen verringert, stellt sie kein adäquates Maß für die Erfassung von tempobedingten Aussprachevariationen dar. Analog zum „am“ wird für diese Wortform die Frikativdauer des einzig konstanten Phons in unmittelbarer Nachbarschaft des getilgten Vokals als zusätzliches Tempomaß herangezogen. Die Lateraldauer scheidet hier aus, da der Lateral je nach Variante ambisilbisch oder im Cluster auftritt. Aussprachevarianten mit einer Elision in der unbetonten Silbe weisen einen labialen Frikativ mit etwa 19 ms kürzerer Dauer auf als solche ohne Elision ($F(1, 148) = 8.76$; $p < .01$). Dabei spielt das Geschlecht keine Rolle.

Mit der Ausnahme der Realisierung von /diː/ als [niː] (8) weisen die drei Wörter „die“, „mir“ und „wie“ nur bis zu fünf einzelne Abweichungen auf, womit sich die Unabhängigkeit der Realisierungsform vom Tempo durch fehlende Varianten erklärt.

Für diese sechs Wortformen, die keine signifikanten Zusammenhänge von PLSR und Aussprachvariation aufweisen, lässt sich erklären, warum ihre Varianten nicht mit Tempounterschieden zusammenfallen: In drei Fällen tritt keine relevante Variation auf („die“, „mir“, „wie“), in einem Fall wurde die Variante ohne Glottalplosiv nicht von dem verwendeten Maß erfasst („am“). Für „vielleicht“ führt die mit der Elision von [ɪ] einhergehende Resilbifizierung zu einem vergleichbaren Tempo, obwohl sich signifikante Segmentverkürzungen ergeben. Auch das Wort „wäre“ weist eine einsilbige Aussprachevariante auf. Der Wegfall der unbetonten Silbe führt aufgrund der Berechnung der PLSR nicht automatisch zur Tempoerhöhung. Ungeklärt bleibt, ob die einsilbigen Varianten der Wortformen typischerweise in einem schneller gesprochenen Kontext auftreten sowie die eigentliche Ursache, warum manche Wortformen keine relevanten Variationen zeigen.

[9] Hierbei sei angemerkt, dass die Entscheidung darüber, ob der Lateral dem Silbenkern entspricht, einzig auf der Beurteilung des Autors nach Inspektion der jeweiligen Fälle beruht und nicht über ein Perzeptionsexperiment verifiziert wurde, und damit zwar einheitlich, aber nicht repräsentativ erfolgte.

11.4.2 Wörter mit tempobedingten Elisionen

An dieser Stelle soll anhand der Wörter „machen“ und „also“ der Zusammenhang zwischen Tempo und Wortrealisierung detaillierter dargestellt werden. Beide Wörter sind zweisilbig und weisen deshalb auch potentiell größere Variabilität in ihren Varianten auf als einsilbige Wortformen. Mit 399 Fällen tritt „also“ besonders häufig auf. Bei „machen“ sind die Realisierungen der unbetonten Silbe von Interesse.

Für „also“ ergeben sich drei häufigere Varianten zur kanonischen Aussprache (67): [alzoː] (189), sowie [azoː] (66) und [ʔazoː] (16). Damit kann das Auslassen eines Glottalverschlusses als Norm angesehen werden. Jede dieser drei Varianten wird mit signifikant höherem Tempo realisiert, [alzoː] etwa 0,7 $^{\text{Sil}'}/_{\text{s}}$ schneller ($t(306) = 2.67$; $p < .01$),[10] [azoː] 2,0 $^{\text{Sil}'}/_{\text{s}}$ ($t(306) = 3.77$; $p < .001$) und [ʔazoː] 1,6 $^{\text{Sil}'}/_{\text{s}}$ schneller ($t(306) = 4.76$; $p < .001$). Einflüsse der Sprecheridentität ergeben sich nicht. Ergebnisse von Post-Hoc Vergleichen zeigen, dass sich einzig [azoː] und [ʔazoː] nicht signifikant in ihrem Tempo voneinander unterscheiden ($p = .84$), was in Anbetracht der wenigen Werte für [ʔazoː] und dem geringen mittleren Unterschied auch nicht zu erwarten war. Damit weisen die beiden häufigeren Varianten nicht nur ein signifikant höheres Tempo gegenüber der kanonischen Realisierung auf, sondern die stärkere Abweichung von [azoː] wird auch noch signifikant schneller realisiert als [alzoː].

Auch bei dem Wort „machen“ ist die kanonische Realisierung nicht die häufigste (49). Relevante Varianten weisen durchweg eine Schwa-Tilgung zugunsten eines silbischen Nasals auf. Dieser ist entweder kanonisch alveolar (31), velar (63), oder labial (18). Hier spielt der Kontext eine wichtige Rolle für das Auftreten der jeweiligen Varianten: So folgt [maχn̩][11] in 17 der 31 Fällen eine Pause, 8 mal ein [n] oder [d]. Nach 16 der 18 labialen Nasale folgen auch labiale Konsonanten ([v] und [m]), sowie zwei Pausen. Für die velare Variante ergibt sich kein dominierender Kontext. Hier treten mitunter auch Pausen und labiale Konsonanten auf. Während also [maχn̩] über den Nullkontext als kanonische Realisierung oder regressive Assimilation mit einem folgenden alveolaren Konsonanten erklärt werden kann, und [maχm̩] über regressive Assimilation mit folgenden labialen Konsonanten, führt der direkte phonetische Kontext zu keiner möglichen Ursache, warum [ŋ̩] auftritt. Zwar wird es sich um eine progressive Assimila-

[10] In diesem Kapitel wird **Sprecher** durchweg als Zufallsfaktor berücksichtigt.

[11] In der Annotation des Kielkorpus lautet diese Variante [maxn̩], da [x] nicht von [χ] unterschieden wird.

tion mit dem vorangehenden [χ] handeln, die Systematik des Auftretens bleibt jedoch ungeklärt. Um sprecherspezifische Varianten handelt es sich nicht, da alveolare und velare Nasale von fast allen Sprechern produziert wurden. Die Post-Hoc Vergleiche nach einer klassischen Varianzanalyse zeigen keine signifikanten Unterschiede im Tempo zwischen [maχn̩] und [maχŋ̩] ($p = .89$), während beide Varianten sogar etwas (0,8 $^{\text{Sil}'}/_{\text{s}}$) langsamer sind als die kanonische Aussprache ($t(130) = 2.12$; $p < .05$). Gleichzeitig ergibt sich ein Nebeneffekt für den Zusammenhang von den Varianten und den Sprechern ($F(61, 96) = 1.59$; $p < 0.05$). Wird dieser Effekt durch die Verwendung von z-transformierter PLSR kontrolliert, erweisen sich die beiden Varianten [maxn̩] und [maxŋ̩] von [maxən] ($p = .98$; $p = .73$) oder von einander ($p = .89$) in ihren Tempi nicht verschieden. Dieses Ergebnis ist besonders interessant. Es zeigt, dass eine Elision nicht automatisch eine Tempoerhöhung nach sich ziehen muss. Des Weiteren fällt nicht jede Aussprachevariante mit einem Anstieg in der lokalen Sprechgeschwindigkeit zusammen. Für die Variante [maxm̩] dagegen bestätigt sich das signifikant höhere Tempo von 1,5 $^{\text{Sil}'}/_{\text{s}}$ ($t(130) = 4.04$; $p < .001$).

Auch bei den übrigen häufigen Wörtern sind Abweichungen von einer kanonischen Transkription mit einem signifikanten Anstieg im lokalen Tempo verbunden. Aufgrund der bisherigen Ergebnisse ist zu vermuten, dass vornehmlich vier Prozesse mit steigendem Tempo auftreten, die auch generell Spontansprache charakterisieren:

1. Schwa-Tilgung, wie bei „machen“,
2. Tilgung der unbetonten Silbe, wie in „wäre“,
3. Fehlen des Glottalplosivs („am“),
4. Tilgung eines Konsonanten im Reim („also“),

wobei sich diese Prozesse für „vielleicht“ und „am“ nur anhand von Phondauern und nicht der PLSR nachweisen lassen.

Die weiteren häufigen Wortformen werden auch auf diese Prozesse untersucht, die Ergebnisse aber etwas komprimierter dargestellt: Für Zweisilber tritt die Schwa-Tilgung bei „habe“ ($t(279) = 3,53$; $p < .001$) und „würde“ ($t(197) = 2,36$; $p < .01$) mit signifikant erhöhtem Tempo auf. Insofern fallen bei beiden Wörtern Schwa-Tilgung und Wegfall der unbetonten Silbe zusammen. Bei „Ihnen“ stellt

die Schwa-Tilgung mit 278 von 284 Fällen die Regel dar. Diese Wortform zeichnet sich stattdessen dadurch aus, dass die Elision der unbetonten Silbe – also eigentlich die Tilgung des silbischen Nasals – mit höherem Tempo einher geht ($t(317) = 2,64$; $p < .01$).

Das Fehlen des Glottalplosivs ist zusätzlich zum „am“ bei allen weiteren der hier untersuchten häufigen Wörter mit erhöhtem lokalen Tempo verbunden, wobei dieses Fehlen weit häufiger ist als die kanonische Form (ich[12], uns[13], auch[14], in[15], ein[16], ist[17], es[18], und[19], Ihnen[20]).

Abgesehen von dem Wegfall initialer Glottalplosive betreffen Tilgungen fast ausschließlich Konsonanten im Reim oder Vokale. Ausnahmen bilden wenige Fälle mit silbischem Nasal, wie die 14 Realisierungen des Wortes „den“ als [n̩], wo zusätzlich zum Vokal auch der initiale Plosiv getilgt wurde. Für einsilbige Wortformen, deren Realisierungen einige solcher Tilgungen von Vokal oder Reimkonsonant aufweisen, sind diese Varianten auch signifikant schneller. Hier ist von Interesse, welches Segment dabei elidiert wird, Vokal oder ein Konsonant. Eine Vokaltilgung als jeweils deutlich häufigere Variante ergibt sich für „ich“[21], „ein“[22], „es“[23] und „Ihnen“[24]. (In den Fußnoten befinden sich die Resultate der statistischen Tests, die für diese dominante Variante ein signifikant höheres Tempo gegenüber der vergleichbaren nicht-elidierten Variante aufweisen.) Dem gegenüber sind Elisionen des Konsonanten im Reim für andere Wortformen die dominante Aussprachevariante (auch[25], ist[26], und[27], gut[28], mal[29], nicht[30], noch[31]). Für die übrigen Wörter ergibt sich keine vornehmliche Aussprachevariante mit

[12] $t(1194) = 6,47$; $p < .001$
[13] $t(144) = 5,06$; $p < .001$
[14] $t(286) = 3,25$; $p < .01$
[15] $t(217) = 3,40$; $p < .001$
[16] $t(98) = 4,26$; $p < .001$
[17] $t(149) = 2,22$; $p < .05$
[18] $t(123) = 2,01$; $p < .05$
[19] $t(249) = 3,81$; $p < .001$
[20] $t(149) = 2,22$; $p < .05$
[21] $t(1103) = 2,32$; $p < .05$
[22] $t(193) = 7,58$; $p < .001$
[23] $t(169) = 2,70$; $p < .01$
[24] $t(149) = 2,22$; $p < .05$
[25] $t(255) = 5,30$; $p < .001$
[26] $t(376) = 2,55$; $p < .05$; für die [t]-Elision
[27] $t(435) = 8,42$; $p < .001$
[28] $t(315) = 4,54$; $p < .001$
[29] $t(240) = 6,08$; $p < .001$
[30] $t(255) = 7,78$; $p < .001$
[31] $t(319) = 5,53$; $p < .001$

Tilgung.[32] Eine Systematik, die Konsonanten- gegenüber Vokaltilgung für die verschiedenen Wortformen erklären könnte, ist hier nicht erkennbar.

11.4.3 Wörter mit tempobedingten Reduktionen

Abgesehen von den vier Prozessen, die mit einem Wegfall von Symbolen in der Annotation verbunden sind, werden auch solche Prozesse untersucht, die über eine Änderung eines Transkriptionssymbols erfasst und in dieser Arbeit *Reduktionen* genannt werden. Sie machen einen geringeren Teil der Aussprachevariationen für die hier untersuchten häufigen Wörter aus. Zusätzlich ergeben sich mehr verschiedene Varianten mit geringeren Fallzahlen, da diese segmentalen Reduktionen ihren Ursprung in Assimilationen haben und damit besonders kontextabhängig sind. Ein Beispiel ist der finale Konsonant in „ich“, der je nach Folgewort verschiedenen partiellen Assimilationen unterworfen ist ([ɪz], [ɪʃ], [ɪj]), während die dominante Variante neben der kanonischen Aussprache der silbische Frikativ ist. Im Folgenden werden nur solche Wortformen dargestellt, deren Aussprachvarianten mit Reduktionen über eine angemessene Anzahl von Fällen verfügen (über 5% im Vergleich zur häufigsten Variante): Während „das“ weitgehend kanonisch realisiert wurde (1007 von 1179 Fälle), bildet die Reduktion des Vokals zum Schwa die häufigste Abweichung (86), die allerdings nicht signifikant ist.[33] Für diese Wortform sind nur die verschiedenen Elisionen mit einem signifikanten Tempoanstieg verbunden. Eine Alternative für „dem“ stellen mit [neːm], [deːn] (24 gegenüber 213 kanonischen Realisierungen) zwei Varianten dar.[34] Bei „den“ wurden ein oder auch beide Konsonanten reduziert, und zwar immer zu Nasalen, in der Koda mit anderem Artikulationsort (57 gegenüber 533 Fälle).[35] Die Wortform „es“ ist signifikant schneller, wenn der Vokal als Schwa realisiert wird.[36] Der Vokal von „bis“ wird in seiner häufigsten Variante frikativisch annotiert.[37] Das Wort „dann“ weist mehrere Varianten auf, die wie bei „den“ vor allem die nasale Variation bei den Konsonanten betreffen, wobei durchaus auch der Vokal reduziert wird.[38] Beim „da“[39] und „der“[40] wird in der

[32] Siehe Anhang A.4 für die allgemeinen Ergebnisse der statistischen Tests bezüglich Sprechtempo und Abweichen von der kanonischen Form.
[33] $t(1024) = -0,96$; $p = .339$
[34] $t(209) = 3,87$; $p < .001$
[35] $t(557) = 7,59$; $p < .001$
[36] $t(197) = 3,36$; $p < .001$
[37] $t(332) = 2,49$; $p < .05$
[38] $t(697) = 8,22$; $p < .001$
[39] $t(513) = 4,59$; $p < .001$
[40] $t(350) = 5,19$; $p < .001$

dominanten Variante der Plosiv nasaliert. Außerdem zeigt sich bei „wir“ die Reduktion des Diphthongs [viːɐ] gegenüber [vɐ] als signifikant ($t(865) = 7,33$; $p < .001$).

In der Zusammenfassung bleiben nennenswerte Reduktionen auf Änderungen nasaler Artikulationsorte (machen, dann, dem, den), die Nasalierung des alveolaren stimmhaften Plosiv (den, dem, dann, da, der) sowie Vokalreduktionen zum Schwa (dann, es) beschränkt. Für die hier nicht explizit aufgeführten Wortformen ergeben sich keine dominanten Aussprachevarianten, die Reduktionen beinhalten, sodass zusätzliche statistische Auswertungen zu den in Anhang A.4 dargestellten nicht durchgeführt werden.

11.5 Zusammenfassung und Diskussion

In diesem Kapitel wurden die Auswirkungen von Tempovariation auf die Aussprache über die symbolische Umschrift und de facto anhand häufiger Funktionswörter überprüft. Innerhalb einer Wortform zeigen die Realisierungen mit steigendem Tempo nicht nur *häufiger* Abweichungen von einer kanonischen Aussprache, sondern diese Abweichungen werden auch *stärker*. Unter Einbeziehung des Glottalplosivs in die kanonische Aussprache sind 30 von 35 untersuchten Wörtern von diesen Auswirkungen betroffen, wobei sich nicht jede Realisierungsstufe signifikant in ihrem Tempo von der vorangehenden unterscheidet. Der Zusammenhang zwischen lokalem Tempo und Aussprachevariante gilt für die Anzahl von Elisionen und Reduktionen (auch Assimilationen), sowie für ein Maß, das diese Veränderungen in der Annotation quantitativ kombiniert.

Der signifikante Zusammenhang zwischen Elisionen und Tempo mag einem Automatismus gleichen, da ein fehlendes Segment bei ansonsten gleichen Segmentlängen selbstverständlich eine Silben- und Wortdauerverkürzung nach sich zieht. Obwohl dieser Effekt nicht überraschend ist, bedurfte er doch einer Überprüfung, da er eben nicht obligatorisch ist. Schließlich kann die Elision aus Ökonomiegründen auch durch Segmentdehnungen mit lokal gleichem Tempo einhergehen, wie es in zwei der drei Varianten von „machen“ der Fall ist. Die Überprüfung vom Zusammenhang zwischen Segmentdauer und Elisionen („am“ und „vielleicht“) zeigt, dass sich mit den Elisionen in beiden Fällen auch benachbarte Segmente verkürzen, also die höhere PLSR nicht allein auf die fehlenden Segmente zurückzuführen ist. Die Reduktionen betreffen fast immer konsonantische Segmente, was aber aber über die Transkriptionspraxis erklären lässt, dass bei

den Funktionswörtern Qualitätsänderungen in Vokalen kaum annotiert wurden. Gerade aufgrund dieser Tatsache wurden die spektralen Messungen in Kapitel 9 vorgenommen.

In anderen Untersuchungen wurde bereits gezeigt, dass ein Abweichen von einer kanonischen Aussprache, die über eine breite Transkription erfasst wird, mit allgemein höherem Tempo zusammenfällt (vgl. Kapitel 3.2). Dies bedeutet erst einmal, dass Wörter, die schneller artikuliert werden, auch eher nicht-kanonisch realisiert werden. Die Ergebnisse dieses Kapitels zeigen jedoch, dass der Tempoeffekt auch für Realisierungen derselben Wortform gilt, im vorliegenden Fall also für häufige und damit grundsätzlich bereits schneller gesprochene Wörter. Die typischen Prozesse sind dabei fast ausschließlich Elisionen (Glottalplosiv, Kodakonsonant und Tilgung unbetonter Vokale), die generell Spontan- und Umgangssprache charakterisieren. Wann jedoch ein Vokal oder ein Kodakonsonant getilgt wird, lässt sich mit den vorhandenen Daten genauso wenig systematisieren, wie das Auftreten von Reduktionen. Letztere sind in den Analysen im vorliegenden Kapitel nicht so zahlreich wie die Elisionen, was allerdings in der Entstehung der Annotation begründet sein kann.

12 Zusammenfassung und Ausblick

12.1 Illustration der Ergebnisse am Beispiel des Wortes „vielleicht“

Die Ergebnisse der drei verschiedenen Ansätze zur Erfassung von tempobedingten Aussprachevariationen sollen hier exemplarisch am Beispiel von „vielleicht“ zusammengefasst werden. In den 190 Realisierungen dieses Wortes zeigen sich bei steigendem Tempo sowohl die im letzten Kapitel dargestellte Elision des [ɪ], wie auch spektrale Reduktionen:

88 mal ist /ɪ/ in diesem Wort als Vokal annotiert. Aufgrund der Anzahl wurden nicht-parametrische Tests durchgeführt. Dabei zeigt der Faktor **Geschlecht** keinen Einfluss auf die Verteilung von lokalem Tempo ($U(47,41) = 1161.5$; $p = .10$). Werden die einzelnen Fälle in langsame und schnelle gegenüber dem Median eingeteilt, erweist sich F_2 in Bark als signifikant niedriger in der Gruppe der schnellen ($U(48,40) = 700$; $p < .05$). Für F_1 ergibt sich kein Effekt ($U(48,40) = 872$; $p = .23$). Die Verminderung von F_2 bei erhöhtem Tempo entspricht einer Zentralisierung. Dies ist erst einmal erstaunlich, da in der Regel von höheren F_2-Werten für [l] ausgegangen wird, die z. B. für männliche Sprecher typischerweise bei etwa 1800 Hz liegen. Anhand der Spektrogramme in Abbildungen 12.1 und 12.2 wird deutlich, dass F_2 des Laterals keine eigene Charakteristik aufweist, sondern eine Transition vom [ɪ] zum [a͡ɪ] darstellt. Zugleich visualisiert sich eine Diskontinuität zwischen [ɪ] und [l] für F_1, die auf die schnelle Bewegung der Zungenspitze bis zum Kontakt mit den Alveolen zurückzuführen ist. Der typisch niedrige Wert für F_1 beim [l] beeinflusst damit nicht sichtbar und nicht signifikant die Werte für [ɪ]. Dagegen kann der signifikante Effekt für F_2 als verstärkte Koartikulation mit dem nach [l] folgenden Diphthong interpretiert werden, der entsprechend niedrige Formantwerte zu Beginn aufweist. Im vorliegenden Fall handelt es sich um eine langsamere und eine schnellere Realisierungen des Wortes „vielleicht“ von dem selben männlichen Sprecher. F_2 ist für den Monophthong in der schnelleren Variante niedriger (1732 Hz gegenüber

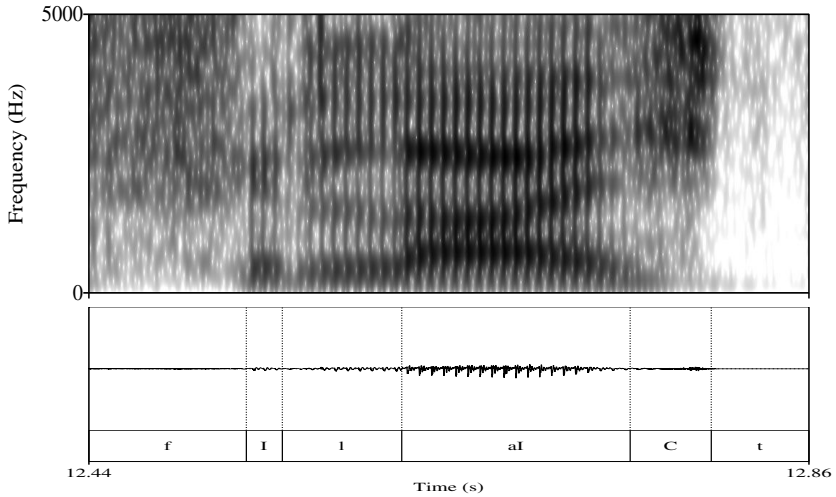

Abbildung 12.1: Oszillogramm und Spektrogramm für das Wort „vielleicht“: langsamere Version

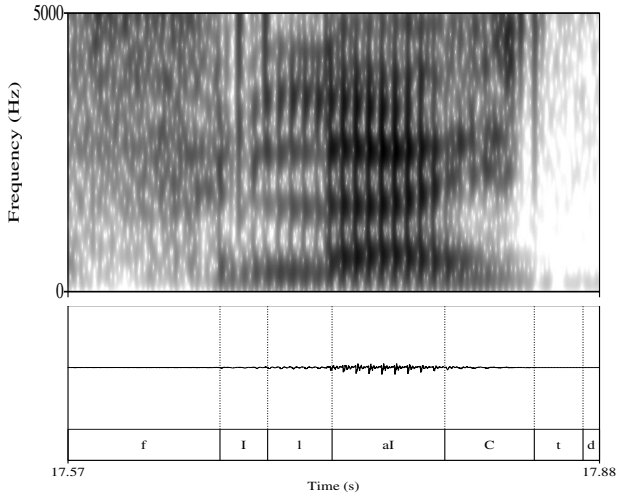

Abbildung 12.2: Oszillogramm und Spektrogramm für das Wort „vielleicht“: schnellere Version

1973 Hz), obwohl beide Dauern fast identisch sind, und auch der Beginn des Diphthongs ist zentralisierter (1572 Hz gegenüber 1385 Hz für F_2).

Die 190 Varianten des /ç/ zeigen höhere Werte im COG für niedrigere Tempi. Langsame Realisierungen (etwa $1/4$ der Daten) weisen um 221 Hz höheres Mittel gegenüber durchschnittlichen auf, schnelle Realisierungen (auch etwa $1/4$ der Daten) ein um 205 Hz geringeres Mittel als durchschnittliche Fälle ($F(1,156) = 36.82$; $p < .001$).

Insofern zeichnet sich das Wort „vielleicht“ durch alle drei in der vorliegenden Arbeit betrachteten Veränderungen bei steigendem lokalen Tempo aus, nämlich durch spektral reduziertere Realisierungen des unbetonten Vokals und post-

vokalischen Frikativs, sowie durch häufigere Elision des Monophthongs.

12.2 Allgemeine Zusammenfassung

Die im Kielkorpus realisierten Aussprachen zeigen signifikante Unterschiede in Abhängigkeit von der Sprechgeschwindigkeit. Diese Unterschiede betreffen nicht nur die Dauern von Segmenten und zeitlichen Informationen zur Lautidentifikation, wie in der Literatur dargestellt, sondern auch spektrale Eigenschaften von Phonen und die lautliche Struktur von Wörtern.

Hier wurde anstatt globaler Raten oder Segmentdauern ein lokales und perzeptiv abgesichertes Maß für die Sprechgeschwindigkeit – die PLSR – verwendet. Der größte Anteil des Einflusses, den das Tempo hat, liegt in der relativen Sprechgeschwindigkeit, die weit stärker variiert als das Tempo zwischen den Sprechern.

Je schneller gesprochen wird, desto stärker sind Monophthonge so koartikuliert, dass insgesamt der Vokalraum schrumpft. Dieser Effekt ist unabhängig von Betonung oder Wortart und damit auch Wortfrequenz. Insofern handelt es sich hierbei um einen echten Tempoeffekt und nicht um ein Resultat veränderten Sprechstils. Aufgrund der sich widersprechenden Ergebnisse aus der Literatur ist es denkbar, dass dies ein fakultatives Phänomen ist. Dennoch kann dieses Ergebnis zusammen mit anderen durchaus über die spezielle Kommunikationssituation der telefonischen Terminabsprache auf sachliche spontane Dialoge verallgemeinert werden. Dass so wenig signifikante Effekte für inter-personelle Unterschiede auftreten, lässt sich auf die geringe Variation zwischen schnellen und langsamen Sprechern zurückführen.

Für Konsonanten wurden stimmlose Frikative analysiert, die mit ansteigendem relativen Tempo spektrale Reduktionen zeigen, die mit verringerter artikulatorischer Engebildung und geringerer Luftstromgeschwindigkeit in Verbindung stehen.

Überraschend ist hierbei das Ergebnis, dass die nachgewiesenen tempo-bezogenen spektralen Effekte für Monophthonge und stimmlose Frikative auch bei unbetonten Silben und in Funktionswörtern auftreten, die bereits kürzer und reduzierter realisiert werden und weniger stark in ihren Dauern vom globalen Tempo beeinflusst sind als solche in betonten Silben.

Auf Wortebene zeigt sich der Einfluss von erhöhter Sprechgeschwindigkeit durch größere Abweichungen von einer kanonischen Aussprache in Form von Elisionen und Reduktionen. Auch hierbei handelt es sich um einen Effekt innerhalb separat untersuchter Gruppen, diesmal für verschiedene Realisierungen eines Wortes. Obwohl fast alle untersuchten Wörter betroffen sind, lässt sich dieses Ergebnis nicht verallgemeinern, da fast ausschließlich häufige Funktionswörter betrachtet wurden.

Für die in dieser Arbeit nachgewiesenen Aussprachevariationen ergibt sich eine einheitliche Tendenz zur Informationsreduktion mit steigendem Tempo. Durch die Einbeziehung der Sprechgeschwindigkeit kommen, trotz Trennung der Daten nach z. B. Betonung, keine invarianten Parameter zum Vorschein. Damit lassen sich auch die Ergebnisse kaum für eine sinnvolle Vorhersage von Aussprache nutzen. Dennoch verringern Tempoinformationen die Streuung der Daten und können diese somit besser beschreiben.

Sowohl Experimente zur Sprachwahrnehmung als auch Evaluierungen automatischer Spracherkenner zeigen, dass sich der Hörer an diese tempobedingten Variationen anpasst, um den Erfolg bei der Sprachverarbeitung zu erhöhen. Die Parameterabschätzungen machen deutlich, dass sich bei den spektralen Messungen zumindest die Monophthonge in Abhängigkeit mit den Tempo so stark verändern, dass von deutlichen Klangänderungen gesprochen werden muss. Durch die hier dargestellten Analysen wird deutlich, dass Tempovariation nicht allein auf Unterschiede zwischen Sprechern beruht, vor allem, da Gesprächspartner vergleichbare globale Tempi zeigen. Vielmehr weist auch alltägliche Sprache einzelner Sprecher große Schwankungen auf. Da weder Sprecheridentität, noch linguistische Informationen zur Betonung und Wortart oder zum wahrscheinlich auftretenden Wort diese Unterschieden in der Aussprache erklären, sind relative Temposchwankungen in Perzeptionsmodelle einzubeziehen.

12.3 Ausblick

Eine Weiterführung der hier vorliegenden Arbeit wäre mit verschiedenen Fragestellungen denkbar. Die hier „tempobedingt“ genannten Aussprachevariationen müssen nicht zwangsläufig ihre Ursache in den Schwankungen der lokalen Sprechgeschwindigkeit haben. Viele bereits als relevant bekannte Faktoren wurden kontrolliert. Jedoch gehen u. a. von van Son und van Santen (2005) davon aus, dass Tempoerhöhung zusammen mit anderen akustischen Reduktionsphä-

nomenen Ausdruck linguistischer oder kontextueller Informationen ist. Dieser These wurde in der vorliegenden Arbeit insofern nachgegangen, dass trotz der Kontrolle bedeutsamer Faktoren sprechtempobedingte Aussprachevariation auftritt. Dennoch kann diese These aber nicht abschließend beurteilt werden, weil noch weitere Faktoren untersucht werden müssten.

Es konnte nicht abschließend geklärt werden, ob sich verschiedene tempobedingte Prozesse, die mittels symbolischer Umschrift erfasst wurden, einer Systematik unterliegen. Da die Anzahl verfügbarer Korpora und Datenbanken gesprochener Sprache zunehmen, wäre es möglich und von Interesse, zu überprüfen, ob sich die hier dargestellten Ergebnisse für häufige und kurze Wortformen systematisieren lassen und auch für niederfrequente Wörter gelten.

Da insbesondere die starke Variation spektraler Parameter unabhängig von dem lokalen Tempo und den kontrollierten Faktoren auffällt, wären anstatt dieser Korpusanalyse mit ihrem großen Datensatz genauerer Einzelanalysen wünschenswert, die spezielle Phänomene, wie etwa Palatalisierung behandeln (vgl. etwa Kohler, 2003a; Moore und Zue, 1985). Auch die Frage nach tempobedingter Resilbifizierung für Wortübergänge wurde hier nicht behandelt. Das Themengebiet der Koartikulation ist insbesondere hinsichtlich Stärke und Auftretensbedingungen tempoinduzierter Veränderungen noch längst nicht umfassend bearbeitet worden. Hierbei wären weitere sich ergänzende akustische und artikulatorische Analysen hilfreich, wie sie bereits bezüglich der Opposition von Lang- und Kurzvokalen durchgeführt wurden (vgl. Kapitel 2.2.3).

Diese starke Variabilität in den Messwerten hat auch ihre Folgen für die Erforschung von Sprachproduktion und Perzeption. Während aktuelle regelbasierte Sprachsynthesesysteme zwar bereits sehr verständlich klingen, wird noch keine zufriedenstellende Natürlichkeit erreicht. Auch die in der vorliegenden Arbeit präsentierten Ergebnisse können aufgrund der großen Datenvarianz nicht für eine gebrauchstüchtige Vorhersage von Aussprachevariation – z. B. für die Sprachsynthese – verwendet werden. Dem gegenüber stellt diese Varianz für eine erfolgreiche automatische Spracherkennung ein geringeres Problem dar. Nicht weitere Forschung, sondern die Implementierung linguistischen Wissens und Weltwissens begrenzt hier die Gebrauchstüchtigkeit. Für die Sprachperzeption ist weniger ungeklärt, *welche* ihm zur Verfügung stehenden Informationen ein Mensch verarbeitet, sondern vielmehr *wie* er sie verarbeitet.

Für die weitere Erforschung der Sprachverarbeitung wäre wünschenswert, die

in der vorliegenden Arbeit dargestellten Effekte auf ihre perzeptive Relevanz zu überprüfen. Dabei erscheint hier weniger bedeutsam, ob sich in uneindeutigen Kontexten Tempoinformationen auf eine linguistische Klassifizierung auswirken, da solche Effekte bereits eindeutig belegt sind und von den jeweiligen Verwechselungsmöglichkeiten und der Stärke von Parameterveränderungen abhängen (vgl. Kapitel 3). Relevanter erscheint hier der noch wenig bearbeitete Bereich der Psycholinguistik, in dem sich mit dem Informationsgehalt von akustischer Variation beschäftig wird: Hilft natürliche Variation bei der Sprachverarbeitung, und hindert unnatürliche Variation diese? Dies wird u. a. mit Hilfe von Reaktionszeitexperimenten erforscht. Es gibt bereits erste Ansätze, Tempoverarbeitung bei der Spracherkennung zu modellieren. Im Rahmen einer Theorie eines episodischen Lexikons (vgl. Kapitel 4.2) müsste überprüft werden, ob sich lokales Tempo als sogenannter Indexfaktor sinnvoller modellieren lässt als intrinsisch, also in den tempo-spektralen Informationen der verschiedenen gespeicherten Exemplare integriert.

Literaturverzeichnis

Abercrombie, D.: *Elements of General Phonetics*. Edinburgh University Press, 1967.

Adams, S. G., Weismer, G. und Kent, R. D.: Speaking rate and speech movement velocity profiles. In: *Journal of Speech and Hearing Research*, Band 36:S. 41–54, 1993.

Adank, P.: *Vowel Normalization: A Perceptual-Acoustic Study of Dutch Vowels*. Ponsen & Looijen, Wageningen, 2003.

Adank, P., Smits, R. und Van Hout, R.: A comparison of vowel normalization procedures for language variation research. In: *Journal of the Acoustical Society of America*, Band 116(5):S. 3099–3107, 2004.

Ainsworth, W. A.: Duration as a cue in the recognition of synthetic vowels. In: *Journal of the Acoustical Society of America*, Band 51(2b):S. 648–651, 1972.

Ainsworth, W. A.: Durational cues in the perception of certain consonants. In: *Proceedings of the British Acoustical Society*. 1973, Band 2, S. 1–4.

Ainsworth, W. A.: The influence of precursive sequences on the perception of synthesized vowels. In: *Language & Speech*, Band 17:S. 103–109, 1974.

Ainsworth, W. A.: Intrinsic and extrinsic factors in vowel judgements. In: Fant, G. und Tatham, M. A. A. (Hg.) *Auditory Analysis and Perception of Speech*, Academic Press, London, S. 103–113. 1975.

Allen, J. S. und Miller, J. L.: Effects of syllable-initial voicing and speaking rate on the temporal characteristics of monosyllabic words. In: *Journal of the Acoustical Society of America*, Band 106(4):S. 2031–2039, 1999.

Anderson, S. und Port, R.: Evidence for syllable structure, stress and juncture from segmental durations. In: *Journal of Phonetics*, Band 22:S. 283–315, 1994.

Assmann, P. F., Neary, T. M. und Hogan, J. T.: Vowel identification: Orthographic, perceptual and acoustic aspects. In: *Journal of the Acoustical Society of America*, Band 71(4):S. 975–989, 1982.

Bard, E. G., Sotillo, C., Kelly, M. L. und Aylett, M. P.: Taking the hit: Leaving some lexical competition to be resolved post-lexically. In: *Language and Cognitive Processes*, Band 16:S. 731–737, 2001.

Batliner, A., Kießling, A., Bürger, C. und Nöth, E.: Filled pauses in spontaneous speech. In: *Proc. of the 13th Intl. Congress of Phonetic Sciences*. 1995, Band 3, S. 472–475.

Becker, T.: *Das Vokalsystem der deutschen Standardsprache*. Peter Lang, Frankfurt/Main, 1998.

Bell, A. und Hooper, J.: Issues and evidence in syllabic phonology. In: Bell, A. und Hooper, J. (Hg.) *Syllables and Segments*, North Holland Publishing, Amsterdam, S. 3–22. 1978.

Bell, A., Gregory, M. L., Brenier, J. M., Jurafsky, D., Ikeno, A. und Girand, C.: Which predictability measures affect content word durations? In: *PMLA*. Estes Park, CO, 2002, S. 1–5.

Bell, A., Jurafsky, D., Fosler-Lussier, E., Girand, C., Gregory, M. und Gildea, D.: Effects of disfluencies, predictability, and utterance position on word form variation in English conversation. In: *Journal of the Acoustical Society of America*, Band 113(2):S. 1001–1024, 2003.

Bell-Berti, F., Regan, S. und Boyle, M.: Final lengthening: Speaking rate effects. In: *Journal of the Acoustical Society of America*, Band 90(4):S. 2311, 1991.

Blache, P. und Meunier, C.: Language as a complex system: The case of phonetic variability. In: *Proceedings of 6. Congreso de Linguística General, Saint-Jacques de Compostelle, Espagne*. 2004.

Bladon, A. und Al-Bamerni, A.: Coarticulation resistance in English /l/. In: *Journal of Phonetics*, Band 4:S. 137–150, 1976.

Boardman, I., Grossberg, S., Myers, C. und Coher, M.: Neural dynamics of perceptual order and context effects for variable-rate speech syllables. In: *Perception & Psychophysics*, Band 61(8):S. 1477–1500, 1999.

Botinis, A., Bannert, R., Fourakis, M. und Pagoni-Tetlow, S.: Crosslinguistic segmental durations and prosodic typology. In: *Speech Prosody, Aix-en-Provence*. 2002, Band 22, S. 183–186.

Boucher, V. J.: A parameter of syllabification for VstopV and relative-timing invariance. In: *Journal of Phonetics*, Band 16:S. 299–326, 1988.

Bradlow, A. R., Torretta, G. M. und Pisoni, D. B.: Intelligibility of normal speech I: Global and fine-grained acoustic-phonetic talker characteristics. In: *Speech Communication*, Band 20:S. 255–272, 1996.

Braunschweiler, N.: Integrated cues of voicing and vowel length in German: A production study. In: *Language & Speech*, Band 40:S. 353–376, 1997.

Brenner-Alsop, E.: Parsing time and rate normalization versus durational contrast. In: *Journal of the Acoustical Society of America*, Band 119(5):S. 3241, 2006.

Browman, C. und Goldstein, L.: Articulatory gestures as phonological units. In: *Phonology*, Band 6:S. 201–251, 1989.

Butterworth, B.: Evidence from pauses in speech. In: Butterworth, B. (Hg.) *Speech and Talk*, Academic Press, New York, Band 1 (Language Production), S. 155–176. 1980.

Byrd, D. und Saltzman, E.: Intragestural dynamics of multiple prosodic boundaries. In: *Journal of Phonetics*, Band 26:S. 173–199, 1998.

Campbell, W. N.: Analog I/O nets for syllable timing. In: *Speech Communication*, Band 9:S. 57–62, 1990.

Campbell, W. N. und Isard, S. D.: Segment durations in a syllable frame. In: *Journal of Phonetics*, Band 19:S. 37–47, 1991.

Carlson, R.: Duration models in use. In: *12. ICPhS, Aix-en-Provence*. 1991, Band 1, S. 243–246.

Carlson, R. und Granström, B.: A search for durational rules in a real-speech data base. In: *Phonetica*, Band 43:S. 140–154, 1986.

Carlson, R. und Granström, B.: Modelling duration for different text materials. In: *1. Eurospeech, Paris*. 1989, Band 2, S. 328–331.

Chait, M., Greenberg, S., Arai, T., Simon, J. und Poeppel, D.: Two time scales in speech segmentation. In: *ISCA Workshop on Plasticity in Speech Perception*. 2005, S. 158.

Chen, M.: Vowel length variation as a function of the voicing of the consonant environment. In: *Phonetica*, Band 22:S. 129–159, 1970.

Cho, T.: *Effects of Prosody on Articulation in English*. University of California, Los Angeles, 2001.

Cho, T. und McQueen, J. M.: Prosodic influences on consonant production in

Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. In: *Journal of Phonetics*, Band 33:S. 121–157, 2005.

Chomsky, N. und Halle, M.: *The Sound Pattern of English*. Harper & Row, New York, 1968.

Christie, W. M.: Some multiple cues for juncture in English. In: *General Linguistics*, Band 17:S. 212–222, 1977.

Chung, G.-Y. und Seneff, S.: A hierarchical duration model for speech recognition based on the ANGIE framework. In: *Speech Communication*, Band 27(2):S. 113–134, 1999.

Cooper, W. E. und Danly, M.: Segmental and temporal aspects of utterance-final lengthening. In: *Phonetica*, Band 38:S. 106–115, 1981.

Crystal, T. und House, A.: Segmental durations in connected speech signals. In: *Journal of the Acoustical Society of America*, Band 83(4):S. 1553–1585, 1988.

Cummins, F.: Some lengthening factors in English speech combine additively at most rates. In: *Journal of the Acoustical Society of America*, Band 105(1):S. 476–480, 1999.

Cutler, A., Dahan, D. und van Donselaar, W. A.: Prosody in the comprehension of spoken language: A literature review. In: *Language & Speech*, Band 40(2):S. 141–202, 1997.

Dalby, J.: *Phonetic Structure of Fast Speech in American English*. Doctoral Dissertation, Indiana University Linguistics Club, 1986.

Dankovičová, J.: The domain of articulation rate variation in Czech. In: *Journal of Phonetics*, Band 25:S. 287–312, 1997.

Dankovičová, J.: Articulation rate variation within the intonation phrase in Czech and English. In: *ICPhS*. San Francisco, 1999, S. 269–272.

Dauer, R.: The reduction of unstressed high vowels in modern Greek. In: *Journal of the International Phonetic Association*, Band 10:S. 17–27, 1981.

Davis, S. und van Summers, W.: Vowel length and closure duration in word-medial VC sequences. In: *Journal of Phonetics*, Band 17:S. 339–353, 1989.

de Jong, K.: The supraglottal articulation of prominence in English: Linguistic stress as localized hyperarticulation. In: *Journal of the Acoustical Society of America*, Band 91(1):S. 491–504, 1995.

de Jong, K.: Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. In: *Journal of Phonetics*, Band 32:S. 493–516, 2004.

de Jong, K. und Zawaydeh, B. A.: Stress, duration, and intonation in Arabic word-level prosody. In: *Journal of Phonetics*, Band 27:S. 3–22, 1999.

Deese, J.: Pauses, prosody and the demands of production in language. In: Dechert, H. W. und Raupach, M. (Hg.) *Temporal Variables in Speech: Studies in Honour of Frieda Goldman-Eisler*, Mouton, The Hague, S. 69–84. 1980.

Delattre, P.: Some factors of vowel duration and their cross-linguistic validity. In: *Journal of the Acoustical Society of America*, Band 34(8):S. 1141–1143, 1962.

Delattre, P., Liberman, A. M., Alvin, M. und Cooper, F. S.: Acoustic loci and transitional cues for consonants. In: *Journal of the Acoustical Society of America*, Band 27(4):S. 769–773, 1955.

Denes, P.: Effect of duration on the perception of voicing. In: *Journal of the Acoustical Society of America*, Band 27(4):S. 761–764, 1955.

Derr, M. A. und Massaro, D. W.: The contribution of vowel duration, *F*0 contour, and frication duration as cues to the /juz/-/jus/ distinction. In: *Perception & Psychophysics*, Band 27(1):S. 51–59, 1980.

Diehl, R. L., Souther, A. F. und Convis, C. L.: Conditions on rate normalization in speech perception. In: *Perception & Psychophysics*, Band 27:S. 435–443, 1980.

Edwards, T. J.: Multiple feature analysis of intervocalic English plosives. In: *Journal of the Acoustical Society of America*, Band 69(2):S. 535–547, 1981.

Engstrand, O.: Articulatory correlates of stress and speaking rate in Swedish VCV utterances. In: *Journal of the Acoustical Society of America*, Band 83(5):S. 1863–1875, 1988.

Engstrand, O. und Krull, D.: Determinants of spectral variation in spontaneous speech. In: *Proc. of Speech Research, Budapest*. 1989, S. 84–87.

Erickson, M. L.: Simultaneous effects on vowel duration in American English: A covariance structure modeling approach. In: *Journal of the Acoustical Society of America*, Band 108(6):S. 2980–2995, 2000.

Essen, O. v.: Sprechtempo als Ausdruck psychischen Geschehens. In: *Zeitschrift für Phonetik*, Band 3:S. 317–341, 1949.

Fant, G.: Speech research in a historical perspective. In: *From Sound to Sense: 50+ Years of Discoveries in Speech Communication*. Boston, 2004, S. 20–41.

Faust, L.: *Variationen von Sprache. Ihre Bedeutung für unser Ohr und für die Sprachtechnologie*. Verlag Dr. Kovač, 1997.

Fischer-Jørgensen, E.: Sound duration and place of articulation. In: *Zeitschrift für Phonetik und Kommunikationsforschung*, Band 17:S. 175–207, 1964.

Fischer-Jørgensen, E.: Intrinsic *F*0 in tense and lax vowels with special reference to German. In: *Phonetica*, Band 47:S. 99–140, 1990.

Fischer-Jørgensen, E. und Jørgensen, H. P.: Close and loose contact ("Anschluß") with special reference to North German. In: *Annual Report of the Institute of Phonetics of the University of Copenhagen (ARIPUC)*, Band 4:S. 43–80, 1969.

Flege, J. E.: Effects of speaking rate on tongue position and velocity of movement in vowel production. In: *Journal of the Acoustical Society of America*, Band 84(3):S. 901–916, 1988.

Fosler-Lussier, E. und Morgan, N.: Effects of speaking rate and word frequency on pronunciations in conversational speech. In: *Speech Communication*, Band 29:S. 137–158, 1999.

Fougeron, C. und Jun, S.-A.: Rate effect on French intonation: Prosodic organisation and phonetic realisation. In: *Journal of Phonetics*, Band 26(1):S. 45–69, 1998.

Fougeron, C. und Keating, P. A.: Articulatory strengthening at edges of prosodic domains. In: *Journal of the Acoustical Society of America*, Band 101(6):S. 3728–3740, 1997.

Fourakis, M.: Tempo, stress, and vowel reduction in American English. In: *Journal of the Acoustical Society of America*, Band 90(4):S. 1816–1827, 1991.

Fowler, C. A.: Coarticulation and theories of extrinsic timing. In: *Journal of Phonetics*, Band 8:S. 113–133, 1980.

Fowler, C. A.: Production and perception of coarticulation among stressed and unstressed vowels. In: *Journal of Speech and Hearing Research*, Band 46:S. 127–149, 1981.

Fowler, C. A.: An event approach to the study of speech perception from a direct-realist perspective. In: *Journal of Phonetics*, Band 14:S. 3–28, 1986.

Fowler, C. A.: Sound-producing sources as objects of perception: Rate normalization and nonspeech perception. In: *Journal of the Acoustical Society of America*, Band 88(3):S. 1236–1249, 1990.

Fowler, C. A. und Housum, J.: Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. In: *Journal of Memory and Language*, Band 26:S. 489–504, 1987.

Fry, D. B.: Duration and intensity as physical correlates of linguistic stress. In: *Journal of the Acoustical Society of America*, Band 27(4):S. 765–768, 1955.

Fujisaki, H. und Kunisaki, O.: Analysis, recognition, and perception of voiceless fricative consonants in japanese. In: *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Band 26(1):S. 21–27, 1978.

Gaitenby, J. H.: The elastic word. In: *Haskins Laboratories Status Report*, Band SR-2:S. 3.1–3.12, 1965.

Gay, T.: Effect of speaking rate on diphthong formant movements. In: *Journal of the Acoustical Society of America*, Band 44(6):S. 1570–1573, 1968.

Gay, T.: Effect of speaking rate on vowel formant movements. In: *Journal of the Acoustical Society of America*, Band 63(1):S. 223–230, 1978.

Gay, T. und Hirose, H.: Effect of speaking rate on labial consonant production: A combined electromyographic/high speed motion picture study. In: *Phonetica*, Band 27:S. 44–56, 1973.

Gendrot, C. und Adda-Decker, M.: Impact of duration on $F1/F2$ formant values of oral vowels: An automatic analysis of large broadcast news corpora in French and German. In: *9th European Conference on Speech Communication and Technology Lisbon*. 2005, S. 2453–2456.

Gentner, D. R.: Timing of skilled motor performance: Tests of the proportional duration model. In: *Psychological Review*, Band 94:S. 255–276, 1987.

Gerstman, L.: Classification of self-normalized vowels. In: *IEEE Transactions on Audio and Electroacoustic*, Band 16:S. 78–80, 1968.

Gibbon, D., Moore, R. und Winski, R. (Hg.): *Handbook of Standards and Resources for Spoken Language Systems*. Mouton de Gruyter, Berlin, 1997.

Goldinger, S. D.: Echoes of echoes? an episodic theory of lexical access. In: *Psychological Review*, Band 105(2):S. 251–279, 1998.

Goldman-Eisler, F.: *Psycholinguistics: Experiments in Spontaneous Speech*. Academic Press, New York, 1968.

Gottfried, T. L. und Beddor, P. S.: Perception of temporal and spectral information in French vowels. In: *Language & Speech*, Band 31:S. 57–75, 1988.

Gottfried, T. L., Miller, J. L. und Payton, P.E.: Effect of speaking rate on the perception of vowels. In: *Phonetica*, Band 47:S. 155–172, 1990.

Green, K. P., Stevens, E. B. und Kuhl, P. K.: Talker continuity and the use of rate information during phonetic perception. In: *Perception & Psychophysics*, Band 88(3):S. 249–260, 1994.

Greenberg, S.: Speaking in shorthand — a syllable-centric perspective for understanding pronunciation variation. In: *Speech Communication*, Band 29:S. 159–176, 1999.

Greenberg, S., Carvey, H., Hitchcock, L. und Chang, S.: The phonetic patterning of spontaneous American English discourse. In: *Proc. ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition, Tokyo*. 2003a.

Greenberg, S., Carvey, H., Hitchcock, L. und Chang, S.: Temporal properties of spontaneous speech—a syllable-centric perspective. In: *Journal of Phonetics*, Band 31:S. 465–485, 2003b.

Grossberg, S.: Resonant neural dynamics of speech perception. In: *Journal of Phonetics*, Band 31:S. 423–445, 2003.

Haggard, M.: Abbreviation of consonants in English pre- and post-vocalic clusters. In: *Journal of Phonetics*, Band 1:S. 9–24, 1973.

Hall, T. A.: *Phonologie. Eine Einführung*. de Gruyter, Berlin, 2000.

Hardcastle, W. J.: Some phonetic and syntactic constraints on lingual coarticulation during /kl/ sequences. In: *Speech Communication*, Band 4:S. 247–263, 1985.

Harrington, J. und Cassidy, S.: Multi-level annotation in the emu speech database management system speech communication. In: *Speech Communication*, Band 33:S. 61–77, 2001.

Harris, K. S.: Cues for the discrimination of American English fricatives in spoken syllables. In: *Language & Speech*, Band 7:S. 1–7, 1958.

Harris, M. S. und Umeda, N.: Effect of speaking mode on temporal factors in speech. In: *Journal of the Acoustical Society of America*, Band 56(3):S. 1016–1018, 1974.

Hawkins, S. und Smith, R.: Polysp: a polysystemic, phonetically-rich approach to speech understanding. In: *Rivista di Linguistica*, Band 13(1):S. 99–188, 2001.

Heid, J. G. G.: Phonetische Variation – Untersuchungen anhand des PhonDat2

Korpus. In: *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der LMU München (FIPKM)*, Band 36:S. 193–368, 1998.

Herrgen, J. und Schmidt, J. E.: Dialektalitätsareale und Dialektabbau. In: Putschke, W., Veith, W. H. und Wiesinger, P. (Hg.) *Dialektgeographie und Dialektologie. Günter Bellmann zum 60. Geburtstag von seinen Schülern und Freunden*, Elwert, Marburg, S. 304–346. 1989.

Hertrich, I. und Ackermannn, H.: Coarticulation in slow speech: Durational and spectral analysis. In: *Language & Speech*, Band 38:S. 157–187, 1995.

Hillenbrand, J. M. und Clark, M. J.: Effects of consonant environment on vowel formant patterns. In: *Journal of the Acoustical Society of America*, Band 109(2):S. 748–763, 2001.

Hirata, Y.: Effect of speaking rate on the vowel length distinction in Japan. In: *Journal of Phonetics*, Band 32:S. 565–589, 2004.

Hirst, D. und Bouzon, C.: The effect of stress and boundaries on segmental duration in a corpus of authentic speech (British English). In: *9th European Conference on Speech Communication and Technology Lisbon*. 2005, S. 29–32.

Hoole, P., Mooshammer, Ch. und Tillmann, H. G.: Kinematic analysis of vowel production in German. In: *Proc. of the 3rd ICSLP, Yokohama*. 1994, S. 53–56.

House, A. S.: On vowel duration in English. In: *Journal of the Acoustical Society of America*, Band 33(9):S. 1174–1178, 1961.

House, A. S. und Fairbanks, G.: The influence of consonant environment upon secondary acoustic characteristics of vowels. In: *Journal of the Acoustical Society of America*, Band 25(1):S. 105–113, 1953.

IPDS: *Kiel Corpus of Spontaneous Speech*. CDROMs 1–3, Kiel, 1995–1997.

Janker, P. M., Pompino-Marschall, B. und Zeynalowa, S.: Variation in the production of german "ein_". In: *Journal of the Acoustical Society of America*, Band 105(2):S. 1399, 1999.

Janse, E.: Word perception in fast speech: artificially time-compressed vs. naturally produced fast speech. In: *Speech Communication*, Band 42:S. 155–173, 2004.

Janse, E., Nooteboom, S. und Quené, H.: Word-level intelligibility of time-compressed speech: prosodic and segmental factors. In: *Speech Communication*, Band 41:S. 287–301, 2003.

Jessen, M.: *Phonetics and Phonology of Tense and Lax Obstruents in German*. John Benjamins, Amsterdam, 1998.

Johnson, K.: Speech perception without speaker normalization: An exemplar model. In: Johnson, K. und Mullennix, J. W. (Hg.) *Talker Variability in Speech Processing*, Academic Press, San Diego, S. 145–166. 1997.

Johnson, K.: Speaker normalization in speech perception. In: Pisoni, D. B. und Remez, R. (Hg.) *The Handbook of Speech Perception*, Blackwell, Oxford, S. 363–389. 2005.

Johnson, K., Ladefoged, P. und Lindau, M.: Individual differences in vowel production. In: *Journal of the Acoustical Society of America*, Band 94(2):S. 701–714, 1993.

Johnson, T. L. und Strange, W.: Perceptual constancy of vowels in rapid speech. In: *Journal of the Acoustical Society of America*, Band 72(6):S. 1761–1770, 1982.

Joos, M.: Acoustic phonetics. In: *Language*, Band 24(2, suppl.):S. 1–136, 1948.

Kaiki, N., Takeda, K. und Sagisaka, Y.: Statistical analysis for segmental duration rules in Japanese speech synthesis. In: *ICSLP, Kobe*. 1990, S. 17–20.

Kato, H., Tsuzaki, M. und Sagisaka, Y.: Measuring temporal compensation in speech perception. In: Sagisaka, Y., Campbell, W. N. und Higuchi, N. (Hg.) *Computing Prosody. Computational Models for Processing Spontaneous Speech*, Springer, New York, S. 251–270. 1997.

Keating, P. A.: Phonetic encoding of prosodic structure. In: Harrington, J. und Tabain, M. (Hg.) *Speech Production: Models, Phonetic Processes, and Techniques*, Psychology Press, New York, S. 167–186. 2006.

Keating, P. A., Cho, T., Fougeron, C. und Hsu, C.-S.: Domain-initial strengthening in four languages. In: Local, J. und Ogden, R., R. Temple (Hg.) *Phonetic Interpretation: Papers in Laboratory Phonology VI*, Cambridge University Press, S. 143–161. 2003.

Kessinger, R. H. und Blumstein, S. E.: Effects of speaking rate on voice-onset time and vowel production: Some implications for perception studies. In: *Journal of Phonetics*, Band 26(2):S. 117–128, 1998.

Kidd, G. R.: Articulatory-rate context effects in phoneme identification. In: *Journal of Experimental Psychology: Human Perception & Performance*, Band 15:S. 736–748, 1989.

Kienast, M.: *Phonetische Veränderungen in emotionaler Sprache*. Shaker, Berlin, 2002.

Klatt, D. H.: Interaction between two factors that influence vowel duration. In: *Journal of the Acoustical Society of America*, Band 54(4):S. 1102–1104, 1973.

Klatt, D. H.: Vowel lengthening is syntactically determined in connected speech. In: *Journal of Phonetics*, Band 3:S. 129–140, 1975.

Klatt, D. H.: Linguistic uses of segmental durations in English: Acoustic and perceptual evidence. In: *Journal of the Acoustical Society of America*, Band 59(5):S. 1208–1221, 1976.

Klatt, D. H.: Synthesis by rule of segmental durations in English sentences. In: Lindblom, B. und Öhman, S. (Hg.) *Frontiers in Speech Communication Research*, Akademic, S. 1174–1178. 1979.

Kohler, K. J.: The production of plosives. In: *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)*, Band 8:S. 30–110, 1977.

Kohler, K. J.: Dimensions in the perception of fortis and lenis plosives. In: *Phonetica*, Band 36:S. 332–343, 1979.

Kohler, K. J.: Stress-timing and speech rate in German. a production model. In: *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)*, Band 20:S. 7–53, 1983.

Kohler, K. J.: Glottal stops and glottalization in German. data and theory of connected speech processes. In: *Phonetica*, Band 51:S. 38–51, 1984a.

Kohler, K. J.: Phonetic explanation in phonology: The feature fortis/lenis. In: *Phonetica*, Band 41:S. 150–174, 1984b.

Kohler, K. J.: Parameters of speech rate perception in German words and sentences: Duration, *F*0 movement, and *F*0 level. In: *Language & Speech*, Band 29:S. 115–139, 1986.

Kohler, K. J.: Segmental reduction in connected speech in German: Phonological facts and phonetic explanations. In: Hardcastle, W. J. und Marchal, A. (Hg.) *Speech Production and Speech Modelling*, Kluwer Academic Publishers, S. 69–92. 1990.

Kohler, K. J.: *Einführung in die Phonetik des Deutschen*. Erich Schmidt, Berlin, zweite Auflage, 1995.

Kohler, K. J.: Domains of temporal control in speech and language: From utterance to segment. In: *Proc. of the 15th ICPhS, Barcelona*. 2003a, S. 7–10.

Kohler, K. J.: Modelling stylistic variation of speech. basic research and speech technology application. In: *Proc. of the 15th ICPhS, Barcelona*. 2003b, S. 223–226.

Koreman, J: Perceived speech rate: The effects of articulation rate and speaking style in spontaneous speech. In: *Journal of the Acoustical Society of America*, Band 119(1):S. 582–596, 2005.

Kowal, S.: *Über die zeitliche Organisation des Sprechens in der Öffentlichkeit*. Huber, Bern, 1991.

Kozhevnikov, V. A. und Christovich, L. A.: *Speech: Articulation and Perception*, Band 30,. translated by the Joint Publications Research Service, 1966, Washington, 1965.

Krause, J. C. und Braida, L. D.: The effects of speaking rate on the intelligibility of speech for various speaking modes (a). In: *Journal of the Acoustical Society of America*, Band 98(5):S. 2982, 1995.

Krech, E.-M., Kurka, E., Stelzig, H., Stock, E., Stötzer, U. und Teske, R. (Hg.): *Großes Wörterbuch der deutschen Aussprache*. Bibliographisches Institut, Leipzig, 1982.

Krull, D., Traunmüller, H. und van Dommelen, W. A.: The effect of local speaking rate on perceived quantity: A comparison between three languages. In: *Proc. of the 15th ICPhS, Barcelona*. 2003, S. 1739–1742.

Kühnert, B.: Some kinematic aspects of alveolar-velar assimilations. In: *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der LMU München (FIPKM)*, Band 31:S. 263–272, 1993.

Kuzlaa, C., Choa, T. und Ernestus, M.: Prosodic strengthening of German fricatives in duration and assimilatory devoicing. In: *Journal of Phonetics*, Band 35:S. 301–320, 2007.

Ladefoged, P. und Maddieson, I.: *The Sounds of the World's Languages*. Blackwell, Oxford, 1996.

Laeufer, Ch.: Patterns of voicing-conditioned vowel duration in French and English. In: *Journal of Phonetics*, Band 20:S. 411–440, 1992.

Laeufer, Ch.: Effects of tempo and stress on German syllable structure. In: *Journal of Linguistics*, Band 33:S. 227–266, 1995.

Lass, N. J.: The significance of intra- and intersentence pause times in perceptual

judgments of oral reading rate. In: *J. Speech & Hearing Research*, Band 13:S. 777–784, 1970.

Laver, J.: *Principles of Phonetics*. Cambridge University Press, 1994.

Lehiste, I.: *Suprasegmentals*. MIT Press, Cambridge, MA., 1970.

Lehiste, I.: The timing of utterances and linguistic boundaries. In: *Journal of the Acoustical Society of America*, Band 51(6b):S. 2018–2024, 1972.

Lehiste, I.: Rhythmic units and syntactic units in production and perception. In: *Journal of the Acoustical Society of America*, Band 54(5):S. 1228–1234, 1973.

Lehiste, I.: Isochrony reconsidered. In: *Journal of Phonetics*, Band 5:S. 253–263, 1977.

Löfquist, A.: Proportional timing in speech motor control. In: *Journal of Phonetics*, Band 19:S. 343–350, 1991.

Liberman, A. M., Delattre, P. C., Gerstman, L. J. und Cooper, F. S.: Tempo of frequency change as a cue for distinguishing classes of speech sounds. In: *Journal of Experimental Psychology*, Band 52:S. 127–137, 1956.

Liberman, A. M., Cooper, F. .S., Shankweiler, D. P. und Studdert-Kennedy, M.: Perception of the speech code. In: *Psychological Review*, Band 74:S. 431–461, 1967.

Lieberman, P.: Some acoustic correlates of word stress in American English. In: *Journal of the Acoustical Society of America*, Band 32(4):S. 451–454, 1959.

Lieberman, P.: Some effects of semantic and grammatical context on the production and perception of speech. In: *Language & Speech*, Band 6:S. 172–187, 1963.

Lindblom, B.: Spectrographic study of vowel reduction. In: *Journal of the Acoustical Society of America*, Band 35(11):S. 1773–1781, 1963.

Lindblom, B.: Explaining phonetic variation: A sketch of the H&H theory. In: Hardcastle, W. J. und Marchal, A. (Hg.) *In Speech Production and Speech Modeling*, Kluwer, Dordrecht, S. 403–439. 1990.

Lippmann, R. P.: Speech recognition by machines and humans. In: *Speech Communication*, Band 22:S. 1–15, 1997.

Lisker, L.: Closure duration and the intervocalic voiced-voiceless distinction in english. In: *Language*, Band 33:S. 42–49, 1957.

Lisker, L.: Rapid vs rabid: A catalogue of acoustic features that may cue the voicing distinction. In: *Haskins Laboratories Status Report*, Band SR-54:S. 127–132, 1978.

Lisker, L.: Letter to the editor. In: *Journal of Phonetics*, Band 10:S. 333–334, 1982.

Lisker, L. und Abramson, A. S.: A cross-language study of voicing in initial stops: Acoustical measurements. In: *Word*, Band 20:S. 384–422, 1964.

Lisker, L. und Abramson, A. S.: The voicing dimension: Some experiments in comparative phonetics. In: *Proc. of the 6th Int. Congress of Phonetic Sciences*. 1970, S. 563–567.

Luce, P. A. und Charles-Luce, J.: Contextual effects on vowel duration, closure duration, and the consonant/vowel ratio in speech production. In: *Journal of the Acoustical Society of America*, Band 87(6):S. 1949–1957, 1985.

Maak, A.: Die Beeinflussung der Sonatendauer durch die Nachbarkonsonanten. In: *Zeitschrift für Phonetik, Sprachwissenschaft und Kommunikationsforschung*, Band 7:S. 104–128, 1953.

Mack, M.: Voicing-dependent vowel duration in English and French: Monolingual and bilingual production. In: *Journal of the Acoustical Society of America*, Band 71(1):S. 173–178, 1982.

Maddieson, I.: Phonetic cues to syllabification. In: Fromkin, V. (Hg.) *Phonetic Linguistics. Essays in Honor of Peter Ladefoged*, Academic Press, New York, S. 203–221. 1985.

Massaro, D. W. und Cohen, M. M.: Consonant/vowel ratio: An improbable cue in speech. In: *Perception & Psychophysics*, Band 33(5):S. 501–505, 1983a.

Massaro, D. W. und Cohen, M. M.: Phonological context in speech perception. In: *Perception & Psychophysics*, Band 34(4):S. 338–348, 1983b.

Mauk, C. E.: Consonant dynamics: Rate- and vowel-dependence. In: *15th ICPhS Barcelona*. 2003, S. 1919–1922.

Max, L. und Caruso, A. J.: Acoustic measures of temporal intervals across speaking rates: variability of syllable- and phrase-level relative timing. In: *Journal of Speech, Language, and Hearing Research*, Band 40:S. 1097–1100, 1997.

Menzerath, P. und de Lacerda, A.: *Koartikulation, Steuerung und Lautabgrenzung: Eine experimentelle Untersuchung*. Ferdinand Dümmlers Verlag, Berlin, 1933.

Menzerath, P. und de Oleza, S. J.: *Spanische Lautdauer. Eine experimentelle Untersuchung*. Walter de Gruyter, Berlin, 1928.

Miller, J. D.: Auditory perceptual interpretation of the vowel. In: *Journal of the Acoustical Society of America*, Band 85:S. 2114–2134, 1989.

Miller, J. L.: Effects of speaking rate on segmental distinctions. In: Eimas, P. D. und Miller, J. L. (Hg.) *Perspectives on the Study of Speech*, Erlbaum, S. 39–74. 1981.

Miller, J. L. und Baer, T.: Some effects of speaking rate on the production of / b / and / w /. In: *Journal of the Acoustical Society of America*, Band 73(5):S. 1751–1755, 1983.

Miller, J. L. und Dexter, E. R.: Effects of speaking rate and lexical status on phonetic perception. In: *Journal of Experimental Psychology: Human Perception & Performance*, Band 14:S. 369–378, 1988.

Miller, J. L. und Liberman, A. M.: Some effects of later occurring information on the perception of stop consonant and semivowel. In: *Perception & Psychophysics*, Band 25(6):S. 457–465, 1979.

Miller, J. L. und Volaitis, L. E.: Effect of speaking rate on the perceptual structure of a phonetic category. In: *Perception & Psychophysics*, Band 46(6):S. 505–512, 1989.

Miller, J. L. und Wayland, S. C.: Limits on the limitations of context-conditioned effects in the perception of [bi] and [wi]. In: *Perception & Psychophysics*, Band 54(2):S. 205–210, 1993.

Miller, J. L., Grosjean, F. und Lomanto, C.: Articulation rate and its variability in spontaneous speech. In: *Phonetica*, Band 41:S. 215–225, 1984.

Miller, J. L., Green, K. P. und Reeves, A.: Speaking rate and segments: A look at the relation between speech production and speech perception for the voicing contrast. In: *Phonetica*, Band 43:S. 106–115, 1986.

Moon, S. J. und Lindblom, B.: Interaction between duration, context, and speaking style in English stressed vowels. In: *Journal of the Acoustical Society of America*, Band 96(1):S. 40–55, 1994.

Moore, K. C. und Zue, V. W.: The effect of speech rate on the application of low-level phonological rules in American English. In: *Journal of the Acoustical Society of America*, Band 77(S1):S. S53, 1985.

Mullennix, J. W.: On the nature of perceptual adjustments to voice. In: Johnson, K. und Mullennix, J. W. (Hg.) *Talker Variability in Speech Processing*, Academic Press, San Diego, S. 67–84. 1997.

Munhall, K., Fowler, C., Hawkins, S. und Saltzman, E.: "compensatory shortening" in monosyllables of spoken English. In: *Journal of Phonetics*, Band 20:S. 225–239, 1992.

Nakatani, L. H., O'Connor, K. D. und Aston, C. H.: Prosodic aspects of American English speech rhythm. In: *Phonetica*, Band 38:S. 84–106, 1981.

Nearey, T. M.: *Phonetic Feature Systems for Vowels*. Indiana University Linguistics Club, 1978.

Nearey, T. M.: Static, dynamic, and relational properties in vowel perception. In: *Journal of the Acoustical Society of America*, Band 85(5):S. 2088–2113, 1989.

Newman, R. S. und Sawusch, J. R.: Perceptual normalization for speaking rate: Effects of temporal distance. In: *Perception & Psychophysics*, Band 58(4):S. 540–560, 1996.

Nittrouer, S., Munhall, K., Kelso, J. A., Tuller, B. und Harris, K. S.: Patterns of interarticulator phasing and their relation to linguistic structure. In: *Journal of the Acoustical Society of America*, Band 84(5):S. 1653–1661, 1988.

Nooteboom, S. G.: *Production and Perception of Vowel Duration*. Doctoral Dissertation, Utrecht, 1972.

Nord, L.: Acoustic studies of vowel reduction in Swedish. In: *Quartely Progress and Status Report, Dept. of Speech Communication, KTH*. 1986, Band 4, S. 19–36.

Nusbaum, H. und Magnuson, J.: Talker normalization: Phonetic constancy as a cognitive process. In: Johnson, K. und Mullennix, J. W. (Hg.) *Talker Variability in Speech Processing*, Academic Press, San Diego, S. 109–132. 1997.

Öhman, S. E. G.: Coarticulation in VCV utterances: Spectrographic measurements. In: *Journal of the Acoustical Society of America*, Band 39(1):S. 151–168, 1966.

Oller, D. K.: The effect of position in utterance on speech segment duration in English. In: *Journal of the Acoustical Society of America*, Band 54(5):S. 1235–1247, 1973.

O'Shaughnessy, D.: A study of French vowel and consonant durations. In: *Journal of Phonetics*, Band 9:S. 385–406, 1981.

Paccia-Cooper, J. und Cooper, W. E.: The processing of phrase structures in speech production. In: Eimas, P. D. und Miller, J. L. (Hg.) *Perspectives on the Study of Speech*, Erlbaum, S. 311–336. 1981.

Paeschke, A.: *Prosodische Analyse emotionaler Sprechweise*. Logos, Berlin, 2003.

Parmenter, C. E. und Treviño, S. N.: The length of the sounds of a Middle Westerner. In: *American Speech*, Band 10:S. 129–133, 1935.

Perkell, J. S. und Klatt, D. H. (Hg.): *Invariance and Variability of Speech Processes*. Lawrence Erlbaum Assoc., Hillsdale, N.J., 1986.

Peterson, G. E. und Barney, H. L.: Control method used in a study of the vowels. In: *Journal of the Acoustical Society of America*, Band 24(2):S. 175–184, 1952.

Peterson, G. E. und Lehiste, I.: Duration of syllable nuclei in English. In: *Journal of the Acoustical Society of America*, Band 32(6):S. 693–703, 1960.

Pfitzinger, H. R.: Local speech rate as a combination of syllable and phone rate. In: *Proc. ICSLP, Sydney*. 1998, Band 3, S. 1087–1090.

Pfitzinger, H. R.: Local speech rate perception in German speech. In: *Proc. of the 14th ICPhS, San Francisco*. 1999, S. 893–896.

Pfitzinger, H. R.: Phonetische Analyse der Sprechgeschwindigkeit. In: *Forschungsberichte des Instituts für Phonetik und Sprachliche Kommunikation der LMU München (FIPKM)*, Band 38:S. 117–264, 2001.

Pfitzinger, H. R.: Intrinsic phone durations are speaker-specific. In: *Proc. ICSLP, Denver*. 2002, Band 2, S. 1113–1116.

Pickett, J. M. und Pollack, I.: Intelligibility of excerpts from fluent speech: effect of rate of utterance and duration of excerpt. In: *Language & Speech*, Band 6:S. 151–164, 1963.

Pike, K.: *Intonation of American English*. University of Michigan Press, Ann Arbor, 1945.

Pinheiro, J. C. und Bates, D. M.: *Mixed-Effects Models in S and S-Plus*. Springer, New York, dritte Auflage, 2001.

Pisoni, D. B.: Some thoughts on "normalization" in speech perception. In: Johnson, K. und Mullennix, J. W. (Hg.) *Talker Variability in Speech Processing*, Academic Press, San Diego, S. 9–32. 1997.

Pisoni, D. B., Carell, T. D. und Gans, S. J.: Perception of the duration of rapid

spectrum changes in speech and nonspeech signals. In: *Perception & Psychophysics*, Band 32(4):S. 141–152, 1982.

Pitermann, M.: Effect of speaking rate and contrastive stress on formant dynamics and vowel perception. In: *Journal of the Acoustical Society of America*, Band 107(6):S. 3425–3437, 2000.

Pols, L. C. W. und van Son, R. J. J. H.: Formant frequencies of Dutch vowels in a text, read at normal and fast rate. In: *Journal of the Acoustical Society of America*, Band 88(4):S. 1683–1693, 1990.

Pols, L. C. W. und van Son, R. J. J. H.: Acoustics and perception of dynamic vowel segments. In: *Speech Communication*, Band 13:S. 135–147, 1993.

Pols, L. C. W., van der Kamp, L. J. T. und Plomp, R.: Perceptual and physical space of vowel sounds. In: *Journal of the Acoustical Society of America*, Band 46(2):S. 458–467, 1969.

Pompino-Marschall, B.: *Die Silbenprosodie. Ein elementarer Aspekt der Wahrnehmung von Sprachrhythmus und Sprechtempo*. Niemeyer, Tübingen, 1990.

Pompino-Marschall, B.: *Einführung in die Phonetik*. de Gruyter, Berlin, erste Auflage, 1995.

Pompino-Marschall, B. und Janker, P. M.: The perception of german syllabic [n]. In: *Journal of the Acoustical Society of America*, Band 105(2):S. 1400, 1999.

Pompino-Marschall, B., Piroth, H.-G., Tilk, K., Hoole, P. und Tillmann, H. G.: Does the closed syllable determine the perception of "momentary tempo". In: *Phonetica*, Band 39:S. 358–367, 1982.

Pompino-Marschall, B., Janker, P. M. und Mooshammer, Ch.: Kinematic and dynamic analysis of German syllables with tense and lax vowels. In: Kröger, B. J., Riek, C. und Sachse, G. (Hg.) *Festschrift Georg Heike*, Hector, Frankfurt/M., S. 161–182. 1998.

Port, R.: *Influence of Speaking Tempo on Duration of Stressed Vowel and Medial Stop in English Trochee Words*. Indiana University, Linguistics Club, Bloomington, 1976.

Port, R.: Linguistic timing factors in combination. In: *Journal of the Acoustical Society of America*, Band 69(1):S. 262–274, 1981.

Port, R. und Dalby, J.: C/V ratio as a cue for voicing in English. In: *Perception & Psychophysics*, Band 32(2):S. 141–152, 1982.

Port, R., Al Ani, S. und Maeda, S.: Temporal compensation and universal phonetics. In: *Phonetics*, Band 37:S. 235–252, 1980.

Posner, M.: Information reduction in analysis of sequential task. In: *Psychological Review*, Band 83:S. 491–503, 1964.

Quené, H.: Modeling of between-speaker and within-speaker variation in spontaneous speech tempo. In: *9th European Conference on Speech Communication and Technology Lisbon*. 2005, S. 2457–2460.

Raphael, J. L.: Preceding vowel duration as a cue to the perception of the voicing characteristic of word-final consonants in American English. In: *Journal of the Acoustical Society of America*, Band 51(4):S. 1296–1303, 1972.

Recasens, D.: Lingual coarticulation. In: Hardcastle, W. J. und Hewlett, N. (Hg.) *Coarticulation*, Cambridge University Press, S. 80–104. 1999.

Repp, B. H. und Mann, V. A.: Perceptual assessment of fricative-stop coarticulation. In: *Journal of the Acoustical Society of America*, Band 67(S1):S. S100, 1980.

Rosen, K. M.: Analysis of speech segment duration with the lognormal distribution: A basis for unification and comparison. In: *Journal of Phonetics*, Band 33:S. 411–426, 2005.

Saberi, K. und Perrott, D. R.: Cognitive restoration of reversed speech. In: *Nature*, Band 398(6730):S. 760, 1999.

Sawusch, J. R. und Newman, R. S.: Perceptual normalization for speaking rate II: Effects of signal discontinuities. In: *Perception & Psychophysics*, Band 62(2):S. 285–300, 2000.

Schmidt, R. A.: A schema theory of discrete motor skill learning. In: *Psychological Review*, Band 82:S. 225–260, 1975.

Schouten, M. E. H. und Pols, L. C. W.: Vowel segments in consonantal contexts: a spectral study of coarticulation—part I. In: *Journal of Phonetics*, Band 7:S. 1–23, 1979.

Schwab, E. C., Sawusch, J. R. und Nusbaum, H. C.: The role of second formant transitions in the stop-semivowel distinction. In: *Perception & Psychophysics*, Band 21(2):S. 121–128, 1981.

Schwartz, M. F.: Identification of speaker sex from isolated voiceless fricatives. In: *Journal of the Acoustical Society of America*, Band 43(5):S. 1178–1179, 1968.

Selkirk, E.: The syllable. In: van der Hulst, H. und Smith, N. (Hg.) *The Structure of Phonological Representations: Part II*, Foris, Dortrecht, S. 337–383. 1982.

Shadle, C. H.: Articulatory-acoustic relationship in fricative consonants. In: Hardcastle, W. J. und Marchal, A. (Hg.) *Speech Production and Speech Modelling*, Kluwer Academic Publishers, S. 187–209. 1990.

Shaiman, S.: Kinematics of compensatory vowel shortening: The effect of speaking rate and coda composition on intra- and inter-articulatory timing. In: *Journal of Phonetics*, Band 29:S. 89–107, 2001.

Shaiman, S., Adams, S. G. und Kimelman, D. Z.: Timing relationships of the upper lip and jaw across changes in speaking rate. In: *Journal of Phonetics*, Band 23:S. 119–128, 1995.

Shattuck-Hufnagel, S. und Turk, A. E.: A prosody tutorial for investigators of auditory sentence processing. In: *ournal of Psycholinguistic Research*, Band 25:S. 193–147, 1996.

Sievers, E.: *Grundzüge der Phonetik zur Einführung in das Studium der Lautlehre der indogermanischen Sprachen*. Breitkopf & Härtel, Leipzig, fünfte Auflage, 1901.

Simpson, A. P.: Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung. In: *Arbeitsberichte des Instituts für Phonetik der Universität Kiel (AIPUK)*, Band 33, 1998.

Sluijter, A. M. C. und van Heuven, V. J.: Effects of focus distribution, pitch accent and lexical stress on the temporal organization of syllables in dutch. In: *Phonetica*, Band 52:S. 71–89, 1995.

Sluijter, A. M. C. und van Heuven, V. J.: Spectral balance as an acoustic correlate of linguistic stress. In: *Journal of the Acoustical Society of America*, Band 100(4):S. 2471–2485, 1996.

Sluijter, A. M. C., van Heuven, V. J. und Pacilly, J. J. A.: Spectral balance as a cue in the perception of linguistic stress. In: *Journal of the Acoustical Society of America*, Band 101(1):S. 503–513, 1997.

Smith, B. L.: Variations in temporal patterns of speech production among speakers of English. In: *Journal of the Acoustical Society of America*, Band 108(5):S. 2438–2442, 2000.

Smith, B. L.: Effects if speaking rate on temporal patterns. In: *Phonetica*, Band 59:S. 232–244, 2002.

Solé, M. J. und Ohala, J. J.: Differentiating between phonetic and phonological processes: The case of nasalization. In: *12th ICPhS Aix-en-Provence*. 1991, S. 110–113.

Stack, J. W., Strange, W., Jenkins, J. J., Clarke, W. D. 3rd und Trent, S. A.: Perceptual invariance of coarticulated vowels over variations in speaking rate. In: *Journal of the Acoustical Society of America*, Band 119(4):S. 2394–2405, 2006.

Stevens, K. N.: Airflow and turbulence noise for fricative and stop consonants: Static considerations. In: *Journal of the Acoustical Society of America*, Band 4B:S. 1180–1192, 1971.

Stevens, K.N. und Blumstein, S.E.: Invariant cues for place of articulation in stop consonants. In: *Journal of the Acoustical Society of America*, Band 64(5):S. 1358–1368, 1978.

Strange, W., Jenkins, J. J. und Johnson, T. L.: Dynamic specification of coarticulated vowels. In: *Journal of the Acoustical Society of America*, Band 74(3):S. 695–705, 1983.

Summerfield, A. Q.: Aerodynamics versus mechanics in the control of voicing onset in consonant-vowel syllables. In: *Speech Perception, Progress Report (Dept. of Psychology, Queen's University, Belfast)*, Band 4:S. 61–72, 1975.

Summerfield, A. Q.: Articulatory rate and perceptual constancy in phonetic perception. In: *Journal of experimental psychology. Human perception and performance*, Band 7:S. 1074–1095, 1981.

Syrdal, A. K. und Gopal, H. S.: A perceptual model of vowel recognition based on the auditory representation of American English vowels. In: *Journal of the Acoustical Society of America*, Band 79(4):S. 1086–1100, 1986.

Thoden, K.: Vorerwähntheit als Einfluss auf die Sprechgeschwindigkeit. *Arbeitsbericht des Graduiertenkollegs „Ökonomie und Komplexität in der Sprache"*. Humboldt-Universität zu Berlin & Universität Potsdam, 2004.

Tillmann, H. G. und Mansell, P.: *Phonetik: Lautsprachliche Zeichen, Sprachsignale und lautsprachlicher Kommunikationsprozeß*. Klett-Cotta, 1980.

Traunmüller, H.: Analytical expression for the tonotopic sensory scale. In: *Journal of the Acoustical Society of America*, Band 88(1):S. 97–100, 1989.

Trouvain, J. und Grice, M.: The effect of tempo on prosodic structure. In: *ICPhS, San Francisco*. 1999, Band 2, S. 1067–1070.

Trouvain, J., Koreman, J., Erriquez, A. und Braun, B.: Articulation rate measures and their relation to phone classification in spontaneous and read German speech. In: *Proc. ISCA-ITR Workshop on Adaptation Methods for Speech Recognition, Sophia-Antipolis*. 2001, S. 155–158.

Tsao, Y.-C. und Weismer, G.: Interspeaker variation in habitual speaking rate: Evidence for a neuromuscular component. In: *Journal of Speech and Hearing Research*, Band 40:S. 858–866, 1997.

Tsao, Y.-C., Weismer, G. und Iqbal, K.: The effect of intertalker speech rate variation on acoustic vowel space. In: *Journal of the Acoustical Society of America*, Band 119(2):S. 1074–1082, 2006.

Tuller, B. und Kelso, J. A. S.: The timing of articulatory gestures. evidence for relational invariants. In: *Journal of the Acoustical Society of America*, Band 76(4):S. 1030–1036, 1984.

Tuller, B. und Kelso, J. A. S.: The production and perception of syllable structure. In: *Journal of Speech and Hearing Research*, Band 34:S. 501–508, 1991.

Tuller, B., Harris, K. und Kelso, J.: Stress and rate: Differential transformations of articulation. In: *Journal of the Acoustical Society of America*, Band 71(6):S. 1534–1543, 1982.

Turk, A. E. und Sawusch, J. R.: The domain of accentual lengthening in American English. In: *Journal of Phonetics*, Band 25:S. 25–41, 1997.

Turk, A. E. und Shattuck-Hufnagel, S.: Word-boundary-related duration patterns in English. In: *Journal of Phonetics*, Band 28:S. 397–440, 2000.

Turner, G. S., Tjaden, K. und Weismer, G.: The influence of speaking rate on vowel space and speech intelligibility for individuals with amyotrophic lateral sclerosis. In: *Journal of Speech and Hearing Research*, Band 38:S. 1001–1003, 1995.

Tyson, N. R.: Applying multiple regression models for predicting word duration in a corpus of spontaneous speech. In: *9th European Conference on Speech Communication and Technology Lisbon*. 2005, S. 2929–2932.

Umeda, N.: Vowel duration in American English. In: *Journal of the Acoustical Society of America*, Band 58(2):S. 434–445, 1975.

Umeda, N.: Consonant duration in American English. In: *Journal of the Acoustical Society of America*, Band 61(3):S. 846–858, 1977.

Utman, J. A.: Effects of local speaking rate context on the perception of voice-

onset time in initial stop consonants. In: *Journal of the Acoustical Society of America*, Band 103(3):S. 1640–1653, 1998.

van Bergem, D. R.: Experimental evidence for a comprehensive theory of vowel reduction. In: *Proc. of 4th. European Conf. on Speech Communication and Technology, Madrid*. 1995, S. 1319–1322.

van Santen, J. P. H.: Contextual effects on vowel duration. In: *Speech Communication*, Band 11:S. 513–546, 1992.

van Son, R. J. J. H. und Pols, L. C. W.: An acoustic description of consonant reduction. In: *Speech Communication*, Band 28:S. 125–140, 1999.

van Son, R. J. J. H. und van Santen, J. P. H.: Duration and spectral balance of intervocalic consonants. In: *Speech Communication*, Band 47:S. 100–123, 2005.

van Summers, W.: Effects of stress and final-consonant voicing on vowel production: Articulatory and acoustic analyses. In: *Journal of Phonetics*, Band 82(3):S. 847–863, 1987.

Venables, W. N. und Ripley, B. D.: *Modern Applied Statistics with S*. Springer, New York, vierte Auflage, 2002.

Vennemann, T.: Skizze der deutschen Wortprosodie. In: *Zeitschrift für Sprachwissenschaft*, Band 10:S. 86–111, 1991.

Verbrugge, R. R. und Shankweiler, D. P.: Prosodic information for vowel identity. In: *Journal of the Acoustical Society of America*, Band 61(S1):S. S39, 1977.

Verbrugge, R. R., Strange, W., Shankweiler, D. P. und Edman, T.: What information enables a listener to map a talker's vowel space? In: *Journal of the Acoustical Society of America*, Band 60(1):S. 198–212, 1976.

Verhoeven, J., de Pauw, G. und Kloots, H.: Speech rate in a pluricentric language: A comparison between Dutch in Belgium and the Netherlands. In: *Language & Speech*, Band 47:S. 297–308, 2004.

Vieregge, W. H.: Phonetische Transkription. Theorie und Praxis der Symbolphonetik. In: *Zeitschrift für Dialektologie und Linguistik*, Band H. 60, Beihefte, 1989.

Vieregge, W. H., Rietveld, A. C. M. und Jansen, C.: A distinctive feature based system for the evaluation of segmental transcription in Dutch. In: *Proc. of the 10th ICPhS, Dordrecht*. 1984, S. 654–659.

Wahlster, W. (Hg.): *Verbmobil: Foundations of Speech-To-Speech Translation*. Springer, Berlin, 2000.

Wayland, S. C., Miller, J. L. und Volaitis, L. E.: Influence of sentential speaking rate on the internal structure of phonetic categories. In: *Journal of the Acoustical Society of America*, Band 95(5):S. 2694–2701, 1994.

Weismer, G. und Fennell, S.: Constancy of (acoustic) relative timing measures in phrase level utterances. In: *Journal of the Acoustical Society of America*, Band 78(1):S. 49–57, 1985.

Weismer, G. und Ingrisano, D.: Phrase-level timing patterns in English: Effects of emphatic stress location and speaking rate. In: *Journal of Speech and Hearing Research*, Band 22:S. 516–533, 1979.

Weiss, B.: Prosodic elements of a political speech and its effects on listeners. In: *10th International Conference on Speech and Computer, Patras*. 2005a, S. 127–130.

Weiss, B.: Variation of local speaking rate in spontaneously produced vowels. In: *16th Conference on Electronic Speech Signal Processing / 15th Czech-German Workshop on Speech Processing Prague*. 2005b, S. 99–106.

Weitkus, K.: *Experimentelle Untersuchung der Laut- und Silbendauer im deutschen Satz*. Dissertation, Bonn, 1931.

Werner, S., Eichner, M., Wolff, M. und Hoffmann, R.: Verwendung eines Sprachmodells zur Modellierung und Synthese von Spontansprache. In: *ESSV*. Karlsruhe, 2003, S. 188–195.

Werner, S., Wolff, M. und Hoffmann, R.: Pronunciation variation modeling for spontaneous speech synthesis. In: *16th Conference on Electronic Speech Signal Processing / 15th Czech-German Workshop on Speech Processing Prague*. Prague, 2005, S. 381–387.

Wightman, C. W., Shattuck-Hufnagel, S., Ostendorf, M. und Price, P.: Segmental duration in the vicinity of prosodic phrase boundaries. In: *Journal of the Acoustical Society of America*, Band 91(3):S. 1707–1717, 1992.

Anhang A: Ergebnisse der statistischen Auswertung

A.1 Unterschiede in den Monophthongen für sprecherbezogene relative Tempovariation

Tabelle A.1 zeigt, in welchen Bedingungen sich bei steigendem Tempo die durchschnittlichen relativen Formantwerte signifikant um einen Schätzwert für 1 SD normalisierter PLSR verändern. Dabei entspricht 1 SD durchschnittlich 1,4 $^{\text{Silben}}/_{\text{s}}$, etwa $^1/_4$ der Spannweite von PLSR für einen Sprecher. Der ermittelte Wert für die Formanten ist besonders von den linguistischen Bedingungen und Geschlecht abhängig und liegt zwischen 80 und 150 Hz.

Die Ergebnisse entstammen einer Regressionsanalyse für jeden Monophthong in jeder Bedingung (betont Inhaltswort, unbetont Inhaltswort, Funktionswort). Von Bedeutung ist wegen der bekannten Streuung nicht der R^2-Wert, sondern das Ergebnis eines F-Tests, welches angibt, ob das Modell die Daten angemessen modelliert ($p < .01$). Trifft dies zu, zeigt ein t-Test, ob die Regressionsgrade eine signifikante Steigung hat ($\alpha = 0.01$). Die folgende Tabelle zeigt die Ergebnisse dieser t-Tests (Freiheitsgerade in Klammern). Das Signifikanzniveau wird mit angegeben (** $\widehat{=}$ $p < .01$; *** $\widehat{=}$ $p < .001$). Ausgelassene Zellen sind nicht signifikant, n. a. bedeutet nicht ausreichend Fälle für eine Analyse vorhanden.

Tabelle A.1: Unterschiede in relativen F1 und F2 für eine relative Tempoerhöhung

Vokal	Betonung	Formant	Wert	t-Wert
aː	betont	F1	$-0,28$	$-8.56(539)^{***}$
	unbetont	F1	$-0,49$	$-14.85(856)^{***}$
	Funktionsw.	F1	$-0,34$	$-11.56(1144)^{***}$
	betont	F2	$+0,19$	$5.19(539)^{***}$
	unbetont	F2	– – –	– – –
	Funktionsw.	F2	$+0,25$	$7.53(1144)^{***}$
a	betont	F1	$-0,14$	$-6.18(1021)^{***}$
	unbetont	F1	$-0,35$	$-7.02(314)^{***}$
	Funktionsw.	F1	$-0,15$	$-7.00(1856)^{***}$
	betont	F2	– – –	– – –
	unbetont	F2	– – –	– – –
	Funktionsw.	F2	$+0,05$	$2.82(1856)^{**}$
eː	betont	F1	$+0,12$	$3.67(667)^{***}$
	unbetont	F1	– – –	– – –
	Funktionsw.	F1	– – –	– – –
	betont	F2	$-0,22$	$-8.02(667)^{***}$
	unbetont	F2	$-0,32$	$-6.55(584)^{***}$
	Funktionsw.	F2	$-0,14$	$-4.05(778)^{***}$
ɛ	betont	F1	$-0,15$	$-5.71(1174)^{***}$
	unbetont	F1	– – –	– – –
	Funktionsw.	F1	– – –	– – –
	betont	F2	– – –	– – –
	unbetont	F2	– – –	– – –
	Funktionsw.	F2	– – –	– – –
iː	betont	F1	$+0,13$	$3.62(839)^{***}$
	unbetont	F1	– – –	– – –
	Funktionsw.	F1	$+0,24$	$7.07(839)^{***}$
	betont	F2	$-0,19$	$-5.30(102)^{***}$
	unbetont	F2	– – –	– – –
	Funktionsw.	F2	– – –	– – –
ɪ	alle	F1/F2	– – –	– – –
oː	betont	F1	$+0,20$	$2.68(244)^{**}$
	unbetont	F1	– – –	– – –
	Funktionsw.	F1	– – –	– – –

Tabelle A.1: Unterschiede in relativen F1 und F2 für eine relative Tempoerhöhung

Vokal	Betonung	Formant	Wert	t-Wert
	betont	F2	$+0,22$	3.54(244)***
	unbetont	F2	– – –	– – –
	Funktionsw.	F2	– – –	– – –
ɔ	betont	F1	$-0,19$	−4.95(576)***
	unbetont	F1	$-0,23$	−2.83(209)***
	Funktionsw.	F1	$-0,14$	−3.43(586)***
	betont	F2	$-0,13$	−3.01(576)**
	unbetont	F2	– – –	– – –
	Funktionsw.	F2	$+0,17$	4.44(586)***
uː	betont	F1	– – –	– – –
	unbetont	F1	– – –	– – –
	Funktionsw.	F1	– – –	– – –
	betont	F2	$+0,25$	4.30(261)***
	unbetont	F2	– – –	– – –
	Funktionsw.	F2	n. a.	n. a. (11)
ʊ	alle	F1/F2	– – –	– – –
øː	betont	F1/F2	– – –	– – –
	unbetont	F1/F2	n. a.	n. a. (0)
	Funktionsw.	F1/F2	n. a.	n. a. (0)
œ	betont	F1	– – –	– – –
	unbetont	F1	n. a.	n. a. (0)
	Funktionsw.	F1	$-0,22$	−3.45(308)***
	betont	F2	– – –	– – –
	unbetont	F2	n. a.	n. a. (0)
	Funktionsw.	F2	– – –	– – –(308)
yː	betont	F1/F2	– – –	– – –
	unbetont	F1/F2	n. a.	n. a. (9)
	Funktionsw.	F1/F2	n. a.	n. a. (1)
ʏ	betont	F1/F2	– – –	– – –
	unbetont	F1/F2	n. a.	n. a. (10)
	Funktionsw.	F1/F2	– – –	– – –
ə	betont	F1	n. a.	n. a. (0)
	unbetont	F1	$-0,09$	−4.36(922)***
	Funktionsw.	F1	– – –	– – –

Tabelle A.1: Unterschiede in relativen F1 und F2 für eine relative Tempoerhöhung

Vokal	Betonung	Formant	Wert	t-Wert
	betont	F2	n. a.	n. a. (0)
	unbetont	F2	− − −	− − −
	Funktionsw.	F2	− − −	− − −
ɐ	betont	F1	n. a.	n. a. (0)
	unbetont	F1	$-0,32$	$-10.45(1220)^{***}$
	Funktionsw.	F1	$-0,29$	$-7.20(577)^{***}$
	betont	F2	n. a.	n. a. (0)
	unbetont	F2	$+0,13$	$4.37(1220)^{***}$
	Funktionsw.	F2	− − −	− − −

A.2 Unterschiede in den Monophthongen zwischen langsamen und schnellen Sprechern

Es gibt nur wenige signifikante Unterschiede in den Formantfrequenzen zwischen der Gruppe langsamer gegenüber schneller Sprecher. Diese sind in Tabelle A.2 eingetragen. Da für diese Fragestellung keine z-Transformation der Daten durchgeführt werden kann, werden die Formantwerte in Bark mit einem gemischten Modell überprüft. Sprecher werden als Zufallsfaktor betrachtet. Unabhängige Variable ist die Gruppe schneller gegenüber langsamer Sprecher ($DF = 1$). Der Freiheitsgrad des Nenners ist jeweils angegeben. Geschlecht und individuelles relatives Tempo bilden Kontrollvariablen. Bei einer Interaktion zwischen der abhängigen Variable und dem Geschlecht, werden die Tests getrennt für Männer und Frauen wiederholt. Der errechnete Wert bezieht sich auf eine Veränderung der Gruppe schneller Sprecher vom Durchschnitt der langsamen. Dabei entsprechen je nach absoluten Formantwerten 100 Hz etwa 0,5 Bark. Das Signifikanzniveau wird nach dem F-Wert angegeben (** $\widehat{=}$ $p < .01$; *** $\widehat{=}$ $p < .001$). Keine Analysen waren möglich für [yː], [uː] in Funktionswörtern.

Tabelle A.2: Unterschiede in F1 und F2 (in Bark) schneller gegenüber langsamen Sprechern; nur signifikante Ergebnisse

Vokal	Betonung	Formant	Wert	F-Wert	DF
aː	unbetont	F1	$-0,20$	21.95***	16
	Funktionsw.	F1	$-0,53$	19.02**	7[1]
a	betont	F1	$-0,08$	9.69**	16
	Funktionsw.	F1	$-0,07$	11.66**	16
eː	betont	F2	$-0,40$	25.70***	16
ɪ	unbetont	F2	$-0,15$	23.08***	16
œ	Funktionsw.	F2	$-0,50$	19.84***	14
ə	unbetont	F2	$-0,33$	11.84**	16
	Funktionsw.	F1	$-0,38$	17.63***	16

[1] Nur für weibliche Sprecher signifikant.

A.3 Statistische Auswertung stimmloser Frikative

Unterschiede in sprechernormierten COG bei variierter PLSR sind für stimmlose Frikative in Tabelle (A.3) eingetragen. Freiheitsgrad des Zählers ist für alle Statistiken 1. Die Freiheitsgrade des Nenners sind jeweils mit angegeben. Die Erwartungswerte haben die Einheit [Hz]. Die Kontrolle der Position innerhalb einer Silbe (Onset, Reim) ist in keinem Fall signifikant.

Tabelle A.3: Tempoabhängige normierte spektrale Balance; nur signifikante Ergebnisse

Frikativ	Betonung	Wert	F-Wert	DF
ʃ	betont	−15,40	4.92*	489
f	betont	−23,45	54.95***	2248
	unbetont	−35,88	29.82***	564
	Funktionsw.	−14,22	10.56**	505
s	betont	−4,29	8.11**	2596
	unbetont	−26,33	79.82***	2690
	Funktionsw.	−17,91	77.16***	3374
ç	betont	−43,51	79.99***	708
	unbetont	−35,88	10.43**	704
	Funktionsw.	−14,22	91.96**	1615
x	betont	−42,29	6.46*	40
	unbetont	−17,28	16.88***	166
	Funktionsw.	−51,12	16,28***	99
χ	betont	−26,04	23.58***	315
	unbetont	−52,89	48.98***	870
	Funktionsw.	−42,25	115.73***	607

A.4 Statistische Auswertung häufiger Wörter auf ihre Transkription

Ergebnisse der F-Tests für die in Kapitel 11 verwendeten Wortformen ohne klitische Zusätze. Signifikanzniveau beträgt $\alpha = 0.5$. Abhängige Variable bildet die **PLSR**, unabhängige ist die Abweichungen von einer kanonischen Form nach Herrgen und Schmidt (1989), Zufallsfaktor ist **Sprecher**. Signifikante Ergebnisse bedeuten hier jedes Mal ein höheres Tempo der Wortrealisierung bei stärkerer Anweichung der Realisierung (vgl. die Erwartungswerte). Das Signifikanzniveau ist mit angegeben (* $\widehat{=}$ $p < .05$; ** $\widehat{=}$ $p < .01$; *** $\widehat{=}$ $p < .001$). Der Freiheitsgrad des Nenners ist in der betreffenden Zeile zu finden, der des Zählers beträgt 1.

Tabelle A.4: Tempoabhängige Unterschiede in der symbolischen Umschrift

Wort	Wert	F-Wert	DF
also	0,37	30.16***	368
ja	1,04	37.60***	1014
ich	0,34	24.57***	1298
wir	0,28	60.04***	916
das	0,40	28.07***	1146
am	–	0.29	399
gut	0,62	17.01***	317
machen	0,40	13.52***	175
uns	0,37	7.28**	172
dem	0,26	6.50*	221
habe	0,38	6.27*	172
mal	0,69	38.05***	244
vielleicht	–	2.46	190
auch	0,67	37.65***	346
dem	0,26	6.50*	221
mir	–	1.91	437
waere	–	0.50	125
bei	0,24	6.11*	269
der	0,40	14.78***	364
Ihnen	0,24	6.50*	249
nicht	0,73	63.26***	274
wie	–	2.08	188

Tabelle A.4: Tempoabhängige Unterschiede in der symbolischen Umschrift

Wort	Wert	F-Wert	DF
bis	0,75	18.91***	358
die	–	2.44	255
in	0,53	19.38***	258
noch	0,79	44.86***	354
da	1,02	26.74***	530
ein	0,61	58.30***	171
ist	0,47	24.02***	463
Sie	0,44	6.42*	290
wuerde	0,30	8.76**	271
dann	0,80	75.00***	758
es	0,22	5.49*	241
und	0,77	179.21***	672

Danksagung

In Bezug auf die Arbeit für diese Dissertation fühle ich mich vielen Menschen zum Dank verpflichtet. Besonders erwähnen möchte ich meinen Betreuer Prof. Dr. Bernd Pompino-Marschall, dem ich an dieser Stelle für seine tatkräftige Unterstützung bei der Themenfindung und Einarbeitung und seine Geduld danken möchte.

Mein herzlicher Dank gilt außerdem meinen Kollegen im Graduiertenkolleg, dem ZAS und den phonetischen Instituten für sehr hilfreiche Gespräche. Insbesondere möchte ich Prof. Dr. Anke Lüdeling, Barbara, Ulrike und Toady danken, die mich immer wieder unterstützt und motiviert haben.

Diese Arbeit enstand im Rahmen des Graduiertenkollegs „Ökonomie und Komplexität in der Sprache“ (GRK 275).

Selbständigkeitserklärung

Hiermit versichere ich an Eides statt, dass ich die vorliegende Dissertation auf der Grundlage der angegebenen Hilfsmittel und Hilfen selbstständig angefertigt habe.

Berlin, den 08.02.2008